

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

ENZO D’ANGELO ARRUDA DUARTE

MODELAGEM DE CONSUMO UTILIZANDO DADOS DE VAZÃO OBTIDOS POR TELEMETRIA EM REDE DE DISTRIBUIÇÃO DE ÁGUA. ESTUDO DE CASO: LAGO NORTE, DISTRITO FEDERAL

Brasília, DF
2024

ENZO D'ANGELO ARRUDA DUARTE

Modelagem de Consumo utilizando Dados de Vazão Obtidos por Telemetria em Rede de Distribuição de Água
Estudo de Caso: Lago Norte, Distrito Federal

Monografia de Projeto Final submetida ao Departamento de Engenharia Civil e Ambiental da Faculdade de Tecnologia da Universidade de Brasília como parte dos requisitos necessários para a obtenção do grau de Engenheiro Ambiental.

Aprovada por:

_____

Dr. Alexandre Kepler Soares (UnB)
(Orientador)

_____

Diogo Fidelis Costa (Caesb)
(Co-orientador)

_____

Prof. Leonardo Zandonadi Moura (UnB)
(Examinador interno)

_____

Davi Navarro de Almeida (Ministério das Cidades)
(Examinador externo)

Brasília, DF
2024

REFERÊNCIA BIBLIOGRÁFICA

ARRUDA DUARTE, E. D. Modelagem de Consumo utilizando Dados de Vazão Obtidos por Telemetria em Rede de Distribuição de Água. Estudo de Caso: Lago Norte, Distrito Federal. Monografia de Projeto Final, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 103 p., 2024.

CESSÃO DE DIREITOS

NOME DO AUTOR: Enzo D'Angelo Arruda Duarte
TÍTULO DA MONOGRAFIA DE PROJETO FINAL: Modelagem de Consumo utilizando Dados de Vazão Obtidos por Telemetria em Rede de Distribuição de Água. Estudo de Caso: Lago Norte, Distrito Federal.
GRAU / ANO: Bacharel em Engenharia Ambiental / 2024.

É concedida à Universidade de Brasília a permissão para reproduzir cópias desta monografia de Projeto Final e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte desta monografia de Projeto Final pode ser reproduzida sem a autorização por escrito do autor.

_____

Enzo D'Angelo Arruda Duarte

enzo.dnglo@gmail.com

Arruda Duarte, Enzo D'Angelo.
AR779 MODELAGEM DE CONSUMO UTILIZANDO DADOS DE VAZÃO OBTIDOS POR TELEMETRIA EM REDE DE DISTRIBUIÇÃO DE ÁGUA. ESTUDO DE CASO: LAGO NORTE, DISTRITO FEDERAL / Enzo D'Angelo Arruda Duarte; orientador Alexandre Kepler Soares; co-orientador Diogo Fidelis Costa. - - Brasília, 2024.
103 p.

Monografia (Graduação – Engenharia Ambiental) - - Universidade de Brasília, 2024.

1. Telemetria. 2. Modelagem Estatística de Consumo. 3. ARIMA. 4. SARIMA. 5. Regressão Linear Multivariada. I. Soares, Alexandre Kepler, orient. II. Costa, Diogo Fidelis, co-orient. III. Título

Dedico este trabalho àqueles que me gritaram “VIVA”, fosse do verbo viver, fosse da intensidade de cada conquista, fosse da memória dos amores que, dia após dia, me trouxeram vida – minha mãe e meu irmão – fosse dos amores que a vida, generosamente, me deu – meus amigos.

Agradeço a todos vocês que hoje também chamo de família.

AGRADECIMENTOS

Agradeço, em primeiro lugar, a minha mãe que sob muito sol, me fez chegar até aqui pela sombra e com água fresca. Seu amor incondicional e cuidado, provenientes d'alma, foram a base para que eu pudesse trilhar meu caminho, viver plenamente e florescer. Ser vida em sua vida foi o maior presente que poderia receber, e cada conquista que celebro, devo a você.

Ao meu irmão, Enri, que diariamente me desafia, seja testando minha paciência, seja provocando minhas habilidades, agradeço a constante parceria. Você sempre esteve de braços abertos quando precisei de acolhimento e me deu o empurrão necessário quando mais precisava de impulso.

Por fim, mas com igual importância, agradeço profundamente aos meus amigos, minha segunda família, que estiveram comigo ao longo de toda essa jornada acadêmica. Sem o apoio, a amizade e o companheirismo de vocês, talvez eu não fosse a pessoa que sou hoje.

A cada um de vocês, minha gratidão, pois, afinal, somos um mosaico das pessoas que mais amamos.

De olhos abertos, eu observo o movimento na estrada e penso na minha trajetória até aqui, nas inúmeras coisas que meus olhos já viram e eu me percebo sendo uma grande colecionadora de boas memórias... É claro, dentro dessas memórias também existem os equívocos, os deslizes, alguns tombos, precipícios, mas é impressionante que mesmo com esses declínios eu ainda lembro como voar.

(Liniker)

RESUMO

A modelagem de redes de distribuição de água é amplamente discutida na literatura científica, buscando aproximar os valores observados com os calculados por meio de modelos estatísticos. O avanço tecnológico permitiu a utilização de sistemas de telemetria e medição inteligente, capazes de fornecer dados detalhados, facilitando o desenvolvimento de modelos de previsão e otimização da gestão hídrica.

Nesse contexto, este trabalho apresenta uma análise do consumo de água na Península Norte, Região Administrativa do Lago Norte, Distrito Federal, onde a Companhia de Saneamento Ambiental do Distrito Federal (CAESB) implementou um sistema avançado de telemetria. O estudo visou modelar o consumo de água utilizando dados de vazão aplicados a três modelos principais: ARIMA, SARIMA e regressão linear multivariada.

Os dados de telemetria foram tratados e modelados em séries temporais para receber os métodos propostos. O modelo ARIMA foi utilizado para capturar as variações do consumo sem sazonalidade explícita, enquanto o modelo SARIMA incorporou padrões sazonais semanais. O modelo de regressão linear multivariada, por sua vez, levou em consideração variáveis explicativas adicionais, como a vazão média, a vazão do dia anterior e de 3 semanas anteriores.

Dessa forma, o trabalho contribui para o desenvolvimento de estratégias mais eficientes de gestão dos recursos hídricos, auxiliando no planejamento e dimensionamento de redes de abastecimento de água em áreas urbanas ao analisar a eficiência de cada um dos modelos na previsão de consumo para a região da Península Norte, concluindo que o modelo SARIMA tem o melhor ajuste dentre os métodos estudados.

Palavras-chave: Água; Consumo; Telemetria; Modelagem; Estatística; ARIMA; SARIMA; Regressão linear multivariada; Sustentabilidade; Gestão hídrica.

LISTA DE FIGURAS

Figura 1.1 - Projeção populacional total do Distrito Federal. 10
Figura 2.1 - Tipos de medidores de vazão. 15
Figura 2.2 - Variações de consumo anual. 23
Figura 2.3 - Variações do consumo diário. 23
Figura 2.4 - Variação do perfil de consumo em função do dia da semana para o setor Itaim Paulista. 30
Figura 3.1 - Localização da Região Administrativa do Lago Norte. 39
Figura 3.2 - Uso e ocupação da RA Lago Norte. 40
Figura 3.3 - Distribuição dos domicílios ocupados segundo o tipo. 41
Figura 3.4 - Abastecimento de água no domicílio. 42
Figura 3.5 - Área de estudo. 43
Figura 3.6 - Configuração hidráulica do sistema de abastecimento de água do Lago Norte. 44
Figura 3.7 - Exemplo de hidrômetro da área de estudo. 45
Figura 3.8 - Dimensões do hidrômetro DIEHL Metering da Hydrus. 46
Figura 4.1 - Autocorrelograma e autocorrelograma parcial da série temporal 55
Figura 4.2 – Série temporal e vazão média dos dados observados. 56
Figura 4.3 - Ajuste ARIMA. 60
Figura 4.4 - Histograma dos resíduos ARIMA. 61
Figura 4.5 - Gráfico de dispersão dos resíduos ARIMA. 61
Figura 4.6 - Previsão ARIMA. 63
Figura 4.7 - Ajuste SARIMA. 65
Figura 4.8 - Histograma dos resíduos SARIMA. 66
Figura 4.9 - Gráfico de dispersão dos resíduos SARIMA. 66
Figura 4.10 - Previsão SARIMA. 68
Figura 4.11 - Matriz de correlação de Pearson para as variáveis independentes - Parte 1. 70
Figura 4.12 - Matriz de correlação de Pearson para as variáveis independentes - Parte 2. 71
Figura 4.13 - Matriz de correlação de Pearson para as variáveis significativas. 72

LISTA DE TABELAS

Tabela 2.1 - Classificação de hidrômetros quanto ao princípio de funcionamento....17
Tabela 2.2 - Classe de hidrômetros e correspondentes valores de Q_{min} e Q_t. 17
Tabela 2.3 - Escore z. 22
Tabela 2.4 - Coeficiente do dia de maior consumo (K_1). 24
Tabela 2.5 - Coeficiente da hora de maior consumo (K_2). 25
Tabela 3.1 - Abastecimento de água no domicílio. 42
Tabela 3.2 - Dados técnicos gerais do hidrômetro DIEHL Metering da Hydrus. 46
Tabela 3.3 - Dimensões do hidrômetro DIEHL Metering da Hydrus. 47
Tabela 4.1 - Valores de volume de consumo entre leituras consecutivas na amostra. 53
Tabela 4.2 - Valores de volume de consumo entre leituras consecutivas na amostra após tratamento. 54
Tabela 4.3 - Valores de volume de consumo médio diários na amostra. 54
Tabela 4.4 - Comparativo de K_1. 57
Tabela 4.5 - Coeficientes modelo ARIMA. 58
Tabela 4.6 - Métricas de avaliação do ajuste ARIMA. 59
Tabela 4.7 - Comparativo modelo ARIMA. 62
Tabela 4.8 - Métricas de avaliação da previsão ARIMA. 62
Tabela 4.9 - Coeficientes modelo SARIMA. 64
Tabela 4.10 - Métricas de avaliação do ajuste SARIMA. 64
Tabela 4.11 - Comparativo modelos SARIMA. 67
Tabela 4.12 - Métricas de avaliação da previsão ARIMA. 67
Tabela 4.13 - Variáveis independentes do modelo de regressão linear múltipla. 69
Tabela 4.14 - Coeficientes estimados para as variáveis explicativas. 71
Tabela 4.15 - Coeficientes estimados para as variáveis significativas. 72
Tabela 4.16 - Métricas de avaliação do modelo de regressão linear multivariada. 73
Tabela 5.1 - Comparativo entre modelos. 76

LISTA DE EQUAÇÕES

Equação 2.1 - Fórmula para cálculo do tamanho da amostra. 21
Equação 2.2 - Coeficiente de consumo de maior vazão diária 22
Equação 2.3 - Coeficiente de consumo de maior vazão horária. 23
Equação 2.4 - Primeira relação da DVGW para coeficientes de consumo. 26
Equação 2.5 - Segunda relação da DVGW para coeficientes de consumo. 26
Equação 2.6 - Fórmula empírica de Tricarico 27
Equação 2.7 - Coeficiente de consumo para 100% de consumo interno. 27
Equação 2.8 - Coeficiente de consumo para 90% de consumo interno. 27
Equação 2.9 - Coeficiente de consumo para 60% de consumo interno. 27
Equação 2.10 - Relação da AWWA para o coeficiente K. 27
Equação 2.11 - Primeira relação da DVGW para o coeficiente K. 27
Equação 2.12 - Segunda relação da DVGW para o coeficiente K. 28
Equação 2.13 - Coeficiente K pela SABESP. 28
Equação 2.14 - Consumo médio efetivo per capita. 29
Equação 2.15 - Consumo *per capita* de água com perdas. 29
Equação 2.16 - Vazão do reservatório até a rede de distribuição. 30
Equação 2.17 - Regressão múltipla. 31
Equação 2.18 - Modelo autorregressivo. 32
Equação 2.19 - Modelo média móvel. 32
Equação 2.20 - Modelo ARMA. 33
Equação 2.21 - Teste de Dickey-Fuller. 33
Equação 2.22 - Diferenciação do modelo ARIMA. 34
Equação 2.23 - Modelo ARIMA. 34
Equação 2.24 - Modelo SARIMA. 34
Equação 2.25 - Método AIC. 36
Equação 2.26 - Teste de Ljung-Box. 36
Equação 2.27 - Erro médio absoluto. 37
Equação 2.28 - Erro médio quadrático. 37
Equação 2.29 - Raiz do erro médio quadrático. 37
Equação 2.30 - Coeficiente de Eficiência de Nash-Sutcliffe. 37
Equação 2.31 - Coeficiente de determinação 38
Equação 2.32 - Erro padrão residual. 38

Equação 4.1 - Formulação de previsão do modelo de regressão linear múltipla. 73

SUMÁRIO

1 INTRODUÇÃO 9
1.1 Objetivos 13
1.1.1 Objetivo Geral 13
1.1.2 Objetivos Específicos 13
2 FUNDAMENTAÇÃO TEÓRICA 14
2.1 Medição de Água 14
2.1.1 Macromedição 14
2.1.2 Micromedição 15
2.2 Telemetria 18
2.3 Tratamento de Dados 20
2.3.1 Identificação e Tratamento de Outliers 20
2.3.2 Amostragem Representativa 21
2.4 Variações de Consumo 22
2.5 Modelagem de Consumo 28
2.5.1 Modelos Aplicados à Dados de Macromedição 28
2.5.2 Modelos Aplicados à Dados de Micro e Macromedição 31
2.5.3 Avaliação e Seleção de Modelos 35
3 METODOLOGIA 39
3.1 Caracterização da Área de Estudo 39
3.2 Modelagem do Consumo de Água 47
3.2.1 Modelo ARIMA 48
3.2.2 Modelo SARIMA 49
3.2.3 Modelo de Regressão Linear Multivariada 49
4 RESULTADOS 51
4.1 Tratamento de Dados 51
4.2 Modelagem de Consumo 55
4.2.1 Série Temporal 55
4.2.2 Modelo ARIMA 58
4.2.3 Modelo SARIMA 64
4.2.4 Modelo de Regressão Linear Multivariada 69
5 CONCLUSÃO E RECOMENDAÇÕES 74
6 BIBLIOGRAFIA 78
7 ANEXOS 81
7.1 Rotina de Tratamento de Dados no RStudio 81
7.2 Rotina ARIMA e SARIMA no RStudio 87
7.3 Rotina Regressão Linear Multivariada no RStudio 97

1 INTRODUÇÃO

A água, como recurso vital para a sustentação da vida, transcende seu papel como mero elemento físico, sendo universalmente reconhecida como um direito fundamental de todo cidadão. Esse reconhecimento foi formalizado em Assembleia Geral das Nações Unidas no dia 28 de julho de 2010 por meio da Resolução 64/292 (ONU, 2010), proclamando o acesso à água limpa e segura e ao saneamento como direitos humanos essenciais para garantia plena a vida. O então secretário, Ban Ki-moon, ressaltou a relevância desse patrimônio ao afirmar que: "A água potável segura e o saneamento adequado são fundamentais para a redução da pobreza, para o desenvolvimento sustentável e para a prossecução de todos e cada um dos Objetivos de Desenvolvimento do Milênio" (ONU, 2010).

Admitir a água como um direito vai além de uma ação simbólica, representa, também, um compromisso com a gestão sustentável desse bem natural. Nesse cenário, a coordenação eficiente dos recursos hídricos é uma preocupação crescente em um mundo cada vez mais impactado pelas mudanças climáticas, crescimento populacional e urbanização acelerada e desordenada. Esses desafios se tornam ainda mais relevantes à medida que o cenário global continua em constante transformação, tal que, de acordo com o Relatório Mundial das Cidades (ONU, 2020), espera-se que a população mundial urbana cresça de 56,2% em 2020 para 60,4% até 2030, intensificando a pressão sobre os sistemas de distribuição de água e demandando uma abordagem integrada e sustentável para o controle hídrico.

No Brasil, a gestão dos recursos hídricos adquire uma importância ainda maior devido à sua vasta e rica hidrografia, aliada à diversidade climática do país e as desigualdades no acesso à água. Regulamentado pela Lei nº 11.445 (Brasil, 2007), e recentemente atualizado pela Lei nº 14.026 (Brasil, 2020), o saneamento básico no Brasil é definido como um conjunto integrado de quatro componentes principais, sendo o abastecimento de água um deles. Essa legislação foi posteriormente reforçada com o novo marco legal do saneamento, sancionado em 2020, que busca universalizar o acesso ao saneamento até 2033, com a meta de garantir que 99% da população tenha acesso à água potável (Brasil, 2020).

Esses avanços legislativos refletem a preocupação crescente com a gestão integralizada dos recursos hídricos e a necessidade de enfrentar os desafios impostos no dia a dia que, mais rotineiramente, resultam em situações de escassez hídrica. A administração adequada da água é fundamental não apenas para garantir o bem-estar

da população, mas também para promover o desenvolvimento econômico e proteger o meio ambiente.

No contexto do Distrito Federal, a situação dos recursos hídricos é especialmente delicada devido ao regime climático da região. As chuvas são concentradas entre setembro e abril, seguidas por um período de seca que ocorre de maio a agosto, durante o qual a umidade relativa do ar frequentemente cai abaixo de 30% (Caesb, 2019). Esse cenário, aliado ao crescimento populacional urbano desorganizado, coloca desafios significativos para a gestão ambiental e influencia na qualidade de vida da população. A projeção populacional para o Distrito Federal, realizada no âmbito do Censo (Instituto Brasileiro de Geografia e Estatística, 2022) e utilizando o método de tendência do crescimento populacional, reflete as inclinações apresentadas na Figura 1.1.

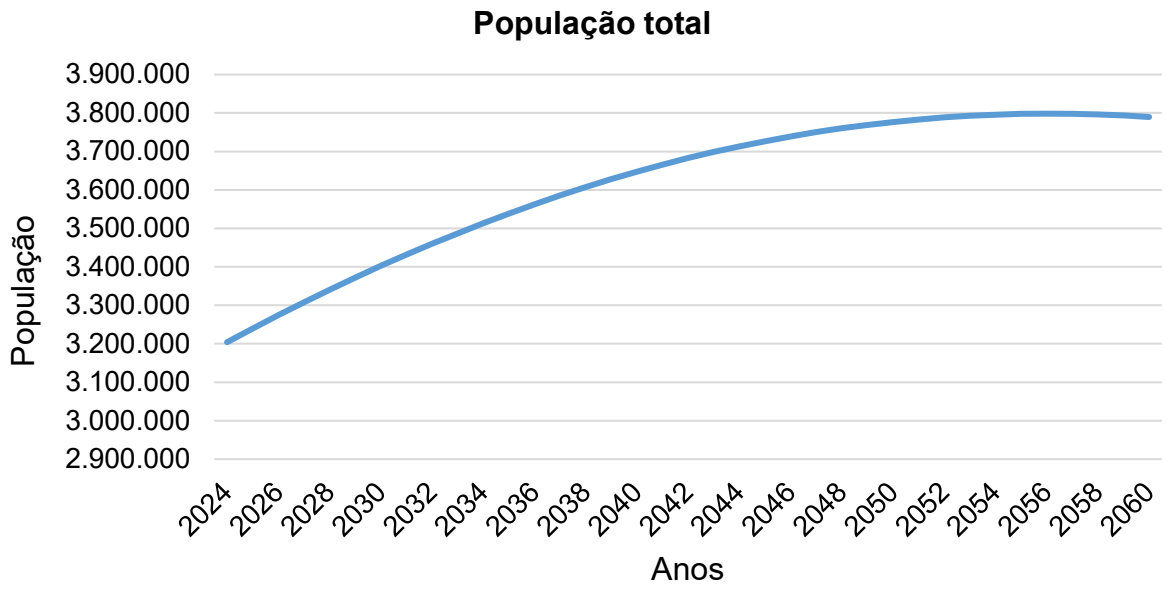


Figura 1.1 - Projeção populacional total do Distrito Federal.

Brasília, embora planejada para ser um modelo de urbanização ordenada, passou por um processo de crescimento desordenado ao longo das décadas, o que a transformou em uma das maiores metrópoles do país. Hoje, consolidada como a terceira maior cidade do Brasil, esse crescimento reflete o acelerado processo de urbanização e os desafios de infraestrutura e planejamento resultantes da expansão descontrolada. Como reportado pelo Jornal O Globo (2023), a região administrativa do Varjão ilustra esse cenário: originada como uma pequena invasão com cerca de 500 famílias, a área cresceu devido à proximidade com o Plano Piloto. Atualmente, o

bairro abriga mais de oito mil moradores, enfrentando problemas graves como a falta de saneamento básico adequado.

Nos últimos anos, portanto, a tecnologia tem desempenhado um papel fundamental na modernização no gerenciamento operacional e hídrico de sistemas de abastecimento de água potável. A telemetria, em particular, tem sido amplamente adotada para o monitoramento em tempo real do consumo de água, permitindo que as empresas e concessionárias de saneamento otimizem a distribuição dos recursos, identifiquem padrões de consumo e detectem vazamentos com mais rapidez e eficiência. Esse tipo de monitoramento também possibilita a implementação de políticas de uso racional da água e aprimora a gestão do consumo, promovendo maior eficiência e adaptabilidade nos sistemas de abastecimento, com base em dados concretos e atualizados.

A literatura acadêmica destaca a relevância de modelagens estatísticas no estudo de previsões de consumo de água. Conforme observado por Costa (2022), a preocupação com o correto dimensionamento das redes de distribuição de água, essencial para assegurar tanto a melhoria operacional quanto o abastecimento adequado, evidencia a importância de estratégias precisas para o entendimento do uso de recursos hídricos. Nesse contexto, a modelagem estatística de curvas de vazão, considerando a variabilidade e a aleatoriedade dos padrões de consumo, tem sido amplamente discutida. Diversas pesquisas visam alinhar os valores históricos observados aos modelados por diferentes métodos, permitindo uma previsão mais acurada do consumo futuro e, consequentemente, uma gestão otimizada da distribuição de água.

A previsão da demanda de água é uma área de estudo que vem ganhando grande relevância, especialmente em regiões onde o consumo varia significativamente ao longo do ano. Modelos estatísticos, como ARIMA (Média Móvel Integrada Autoregressiva) e SARIMA (ARIMA Sazonal), têm se mostrado eficazes na previsão de séries temporais, assim, esses modelos são amplamente utilizados em estudos de previsão de demanda a curto prazo conforme apontado por Anele, Hamam, *et al.* (2017).

Além dos modelos baseados em séries temporais, a regressão linear multivariada oferece uma abordagem complementar, permitindo a inclusão de múltiplas variáveis explicativas no modelo. No caso deste estudo, são consideradas variáveis como a vazão do dia anterior e vazões de dias correspondentes nas

semanas anteriores, além de outros fatores que podem influenciar o consumo de água, como temperatura, precipitação e eventos sazonais. A combinação dessas abordagens estatísticas oferece uma ferramenta poderosa para a previsão de consumo, permitindo a criação de modelos robustos e capazes de capturar a complexidade dos padrões de consumo de água.

No Distrito Federal, a região da Península Norte do Lago Norte se destaca pela implementação pioneira de um sistema de telemetria avançado, que coleta dados de vazão a partir de hidrômetros eletrônicos instalados em cada unidade consumidora. Esse sistema, que envia leituras a cada quatro horas desde 11 de junho de 2019, permite uma visão detalhada e precisa do consumo de água na região, facilitando a identificação de anomalias e a operação e manutenção eficiente das redes de distribuição e seus atributos. A utilização desses dados em estudos acadêmicos abre novas possibilidades para a análise de variações no consumo de água e a modelagem preditiva.

A escolha desta região como estudo de caso se justifica pela sua relevância dentro do contexto distrital, onde a Companhia de Saneamento Ambiental do Distrito Federal (Caesb) tem investido significativamente seu trabalho para com o monitoramento e a gestão da água. O presente estudo visa, então, utilizar esses dados para testar e validar modelos estatísticos preditivos, avaliando sua eficácia na previsão de variações no consumo e sua aplicabilidade para otimizar a operação na região.

Portanto, este projeto final busca contribuir para a literatura existente ao explorar a integração entre tecnologia de medição inteligente e modelagem preditiva, com foco em uma região da Península Norte do Distrito Federal. A hipótese central deste trabalho é que a utilização de dados de telemetria pode fornecer uma base sólida para a previsão de consumo de água, permitindo uma gestão mais eficiente e sustentável dos recursos hídricos. Além disso, o uso de modelos estatísticos como ARIMA, SARIMA e regressão linear multivariada oferece uma abordagem metodológica robusta, capaz de capturar a complexidade das variações de consumo e auxiliar no planejamento de longo prazo das operações de distribuição, permitindo a utilização de tais métodos para o dimensionamento de redes de abastecimento em projetos futuros ou em projetos em áreas com características semelhantes.

1.1 Objetivos

1.1.1 Objetivo Geral

Modelar o consumo de água, refinar e analisar a adequação de métodos preditivos à série temporal de vazão na Península Norte, situada na Região Administrativa do Lago Norte, empregando dados provenientes de sistemas de telemetria.

1.1.2 Objetivos Específicos

- Investigar a série temporal de consumo, abrangendo o valor médio diário de consumo;
- Avaliar criticamente a metodologia de modelagem aplicada, visando aprimoramentos para previsão de consumo, otimização operacional e dimensionamento eficiente da rede em futuros projetos;
- Efetuar uma análise do ajuste e da previsão de consumo em relação aos modelos adotados, com o intuito de validar e fornecer insumos para dimensionamentos em áreas semelhantes.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Medição de Água

A medição de vazão de água desempenha um papel fundamental na gestão eficiente dos recursos hídricos, fornecendo dados importantes para a operação e o monitoramento de sistemas de abastecimento e distribuição. Este processo envolve a quantificação de água que passa por um ponto específico ao longo do tempo, sendo essencial para avaliar sistemas hidráulicos, identificar vazamentos, otimizar o consumo e garantir a conformidade com normas e regulamentos ambientais. A precisão e confiabilidade das medições de vazão são cruciais para tomadas de decisão informadas, promovendo a eficiência no uso da água e contribuindo para a sustentabilidade dos recursos hídricos.

Dessa forma, a medição de vazão abrange duas abordagens específicas, a macro e a micromedição.

2.1.1 Macromedição

A macromedição refere-se à avaliação em larga escala, geralmente realizada em pontos estratégicos de uma rede de distribuição de água, como logo a jusante do reservatório de distribuição. Esses pontos de macromedição fornecem dados representativos sobre o desempenho geral do sistema, permitindo avaliações abrangentes do consumo, identificação de perdas e monitoramento operacional por meio da quantidade volumétrica que passa pela tubulação por unidade de tempo. Segundo Filho e Dall'Aglio (2008), o macromedidor de vazão mais empregado pelas concessionárias brasileiras é do tipo eletromagnético, seguido pelos tipos *Annubar* e ultrassônico.

Adicionalmente, na dissertação de Filho (2007), são destacados diversos tipos de macromedidores de vazão amplamente empregados em sistemas de abastecimento de água. Entre eles, mencionam-se o tipo *Venturi*, tipo turbina, tipo *Pitot*, tipo eletromagnético (classificados em bobina de indução externa, de inserção com um sensor e de inserção com múltiplos sensores), além de medidores ultrassônicos, seja de tempo de trânsito ou de efeito Doppler.

Os macromedidores podem ser categorizados em medidores indiretos, diretos e especiais (USP, 2017). Essa classificação abrange uma variedade de dispositivos projetados para medir a vazão em diferentes contextos. Os medidores indiretos envolvem métodos que inferem a vazão com base em princípios distintos, enquanto

os diretos realizam a medição de maneira mais imediata e por meio do fluxo d'água. Além disso, os medidores especiais são projetados para atender a requisitos específicos em situações particulares. A Figura 2.1 oferece uma visão mais aprofundada dessa classificação, destacando os medidores citados anteriormente associados a cada categoria.

O medidor de vazão eletromagnético, em especial, opera com base no princípio estabelecido por Michael Faraday, que demonstra a geração de corrente elétrica em um campo magnético. Neste dispositivo, duas bobinas de campo são instaladas, criando um campo magnético constante em toda a seção transversal do conduto e paralelamente, dois eletrodos são posicionados para captar a voltagem dentro do tubo. Quando o fluxo é iniciado, as partículas carregadas positiva e negativamente são separadas pelo campo magnético induzido pelas bobinas, gerando uma voltagem detectada e medida pelos eletrodos, tal que quanto maior a velocidade de fluxo, maior a separação das partículas carregadas e maior a voltagem entre os eletrodos.

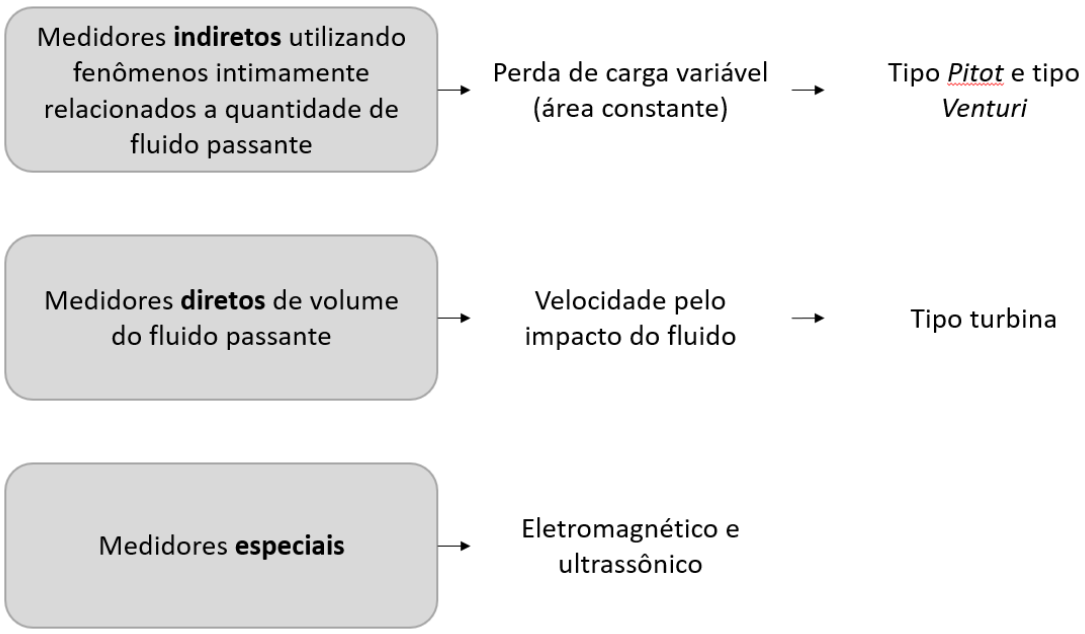


Figura 2.1 - Tipos de medidores de vazão.

Fonte: Adaptado (USP, 2017).

2.1.2 Micromedição

Por sua vez, a micromedição se concentra em uma escala mais localizada, proporcionando uma análise detalhada dos volumes de água consumidos em

residências, comércios, indústrias e locais públicos. A micromedição, portanto, é fundamental para a quantificação precisa do consumo de água pelos usuários finais, sendo crucial para fins de faturamento (Alegre, Coelho, *et al.*, 2005).

Para efetuar a medição precisa da vazão consumida, dispositivos especializados, conhecidos como hidrômetros, são estrategicamente instalados em pontos específicos. Esses hidrômetros têm como principal função realizar comparações entre as leituras registradas durante o período analisado e as do período anterior. Essa comparação resulta na diferença entre esses dois conjuntos de leituras, refletindo o consumo registrado. Conforme definido pela NBR 8009 (ABNT, 1997), o hidrômetro ou medidor de água, é conceituado como um instrumento destinado a indicar e totalizar, de forma contínua, o volume de água que passa por ele.

Os medidores de água consistem em três partes essenciais: a câmara de medição, responsável por se movimentar de acordo com o fluxo de água; um registrador, que converte o movimento gerado pelo fluxo d'água em uma medida de consumo; e, por fim, o mecanismo que transfere o movimento da água da câmara até o registrador (Mendonça *apud* Silva, 2008).

Conforme destacado por Costa (2022), os hidrômetros são classificados com base em seus princípios de funcionamento, categorizando-se como velocimétricos, volumétricos, ultrassônicos e eletromagnéticos. Além disso, de acordo com as diretrizes do INMETRO (2022), é crucial que esses medidores sejam projetados e fabricados de maneira a garantir que seus erros de indicação estejam dentro dos limites máximos permitidos, podendo ser categorizados sob a classe de exatidão 1 ou classe de exatidão 2.

A Tabela 2.1 e a Tabela 2.2, apresentam a classificação dos hidrômetros com base no princípio de funcionamento e na exatidão, respectivamente. Essa categorização é essencial para compreender a diversidade de hidrômetros disponíveis, permitindo a escolha do instrumento mais adequado para atender às demandas específicas de medição de vazão de água em diferentes contextos.

Tabela 2.1 - Classificação de hidrômetros quanto ao princípio de funcionamento.

Classificação	Princípio de funcionamento
Velocimétrico	Parte móvel (palheta, hélice, turbina) que é acionada diretamente pelo fluxo de água e transmite o movimento para o dispositivo indicador.
Volumétrico	Câmara de enchimento cilíndrica (com volume pré-definido) com uma peça móvel (pistão ou disco) que se movimenta a cada enchimento completo e transmite o movimento para dispositivo indicador
Ultrassônico	Baseado na velocidade de propagação de ondas sonoras. Contam com sensores que emitem e recém sinais ultrassônicos.
Eletromagnético	Considera a lei de indução eletromagnética para identificação do volume de água que passa pelo equipamento.

Fonte: Adaptada (Costa, 2022).

Tabela 2.2 - Classe de hidrômetros e correspondentes valores de Q_{min} e Q_t.

Classe	Erro máximo Admissível	
	Campo superior de medição ($Q_2 \leq Q \leq Q_4$)	**Campo inferior de medição ($Q_1 \leq Q \leq Q_2$)**
1	± 1% em temperatura da água entre 0,1ºC e 30ºC	± 3%, independente da faixa de temperatura da água
	± 2% em temperaturas da água acima de 30ºC	
2	± 2% em temperaturas da água entre 0,1ºC e 30ºC	± 5%, independente da faixa de temperatura da água
	± 3% em temperaturas da água acima de 30ºC	

Fonte: Adaptada (INMETRO, 2022).

A classificação do INMETRO (2022) leva em consideração a vazão de sobrecarga (Q_4), que é a maior vazão na qual o medidor deve funcionar por um curto espaço de tempo dentro do seu erro máximo admissível. Ao mesmo tempo, deve

manter seu desempenho metrológico em operações subsequentes dentro de suas condições de utilização. Tal que, também são considerados a vazão de transição (Q_2), que é a razão situada entre a vazão permanente Q_3 e a vazão mínima Q_1. Essa medida divide a faixa de vazão operacional em dois campos, campo superior de medição e campo inferior de medição, cada qual caracterizado pelo seu próprio erro máximo admissível. Por fim, a vazão mínima (Q_1) representa a menor vazão na qual o medidor deve funcionar dentro do erro máximo admissível.

2.2 Telemetria

Atualmente, na maioria das residências brasileiras, a medição do consumo de água ocorre apenas uma vez por mês, não contando com sistemas de avaliação em tempo real. Diante dessa realidade, surge a telemetria, uma abordagem que visa monitorar a rede e as variações de consumo em tempo real por meio da aplicação do conceito de internet das coisas (IoT) aos medidores de vazão inteligentes. Essa tecnologia possibilita a análise de perdas de água, a revisão de procedimentos operacionais e a promoção do consumo consciente pela sociedade (Lima e Navas, 2012). Portanto, o sistema de telemetria recebe as medições coletadas por pulsos dos medidores eletrônicos, transmitindo-as ao centro de processamento de dados da concessionária por meio de recursos de comunicação acoplados aos dispositivos de medição.

Dessa forma, Lima e Navas (2012) estratificam os sistemas de medição remota em três segmentos distintos: a infraestrutura de medição, incumbida da coleta e organização dos dados de vazão coletados, a rede de comunicação, responsável por transferir os dados do ponto de coleta até a central de tratamento de dados, e a infraestrutura de tratamento e gestão de dados, composta pelos servidores e banco de dados projetados para proporcionar uma visão clara e estratégica dos parâmetros monitorados.

A infraestrutura de medição desempenha um papel fundamental na obtenção e transmissão eficiente de dados relacionados ao consumo de água. Essa infraestrutura é composta pelos hidrômetros, responsáveis pela medição direta do volume de água consumido, e pelos dispositivos acoplados a eles que promovem a captura e transmissão desses dados. Esses dispositivos incluem o leitor de pulsos, que registra as leituras do hidrômetro, e o dispositivo de telemetria e telecomando, alimentado por bateria, que transmite os dados para a central de tratamento.

Além disso, para otimizar a comunicação entre os dispositivos e a central, são empregados repetidores de sinal. Esses repetidores desempenham a função de amplificar e estender o alcance da transmissão dos dados, quando necessário, assegurando que as informações coletadas pelos hidrômetros sejam eficientemente transmitidas até o concentrador. O concentrador, por sua vez, recebe e consolida os dados provenientes de diversos dispositivos, encaminhando-os para o sistema de supervisão (Costa, 2022).

Segundo Costa (2022), a rede de comunicação, por sua vez, desempenha um papel fundamental na transmissão eficaz de dados. Essa rede é responsável pela comunicação entre os equipamentos de medição e os concentradores, bem como entre os concentradores e os servidores.

No âmbito da comunicação entre equipamentos de medição e concentrador, frequentemente implementada por redes sem fio, a transmissão de radiofrequência é amplamente adotada. Isso é realizado por meio de transmissores, encarregados de transmitir os dados, e receptores, responsáveis por decodificar as informações recebidas. As redes de transmissão utilizam tecnologias como *Wi-Fi* e *Bluetooth*, sendo notáveis o *Zigbee* e o *LoRaWan*, particularmente destacados em sistemas de telemetria.

Quanto à transmissão de dados que ocorre entre os concentradores e os servidores, diferentes opções são empregadas, incluindo redes de banda larga, GPRS, 3G e 4G.

Finalmente, a infraestrutura de tratamento e gestão de dados desempenha um papel central na operação do sistema, abrangendo servidores dedicados que atuam como fonte de informação, visualização e supervisão dos dados de vazão provenientes tanto dos macromedidores quanto dos micromedidores. Essa parte crítica da infraestrutura armazena os dados em um sistema de gerenciamento especializado, proporcionando uma visão abrangente e detalhada do comportamento do sistema de medição.

O sistema de gestão e tratamento de dados pode ser conceituado como servidores web que desempenham um papel crucial ao alimentar um banco de dados destinado à gestão do conhecimento. Sua principal função reside na construção de uma visão clara e objetiva dos elementos envolvidos no processo, incluindo os consumidores, os recursos disponíveis e a corresponsabilidade social no que diz respeito à eficiência energética (Lima e Navas, 2012).

2.3 Tratamento de Dados

Ao analisar dados de consumo de água, o tratamento de dados desempenha um papel fundamental pois, em muitos casos, os registros mostram flutuações repentinas, também chamadas de picos, que podem distorcer a interpretação dos dados. Esses aumentos abruptos no consumo podem ser resultado de problemas nos medidores de água ou mesmo de falhas na transmissão das informações, que se não forem tratados adequadamente, podem afetar a precisão das análises estatísticas realizadas, levando a conclusões equivocadas.

Em situações como essa é essencial assegurar a qualidade dos dados para desenvolver modelos que representem fielmente o padrão de consumo de água, pois informações com irregularidades e grandes variações, também chamadas de *spikes*, dificultam a detecção de tendências e a precisão das análises estatísticas, o que pode resultar em decisões pouco confiáveis.

2.3.1 Identificação e Tratamento de Outliers

Desse modo, o método do intervalo interquartil (IQR) é amplamente utilizado para a detecção e remoção de *outliers* e *spikes* em conjuntos de dados (Pastega, 2023). O IQR mede a dispersão dos dados, focando no intervalo entre o primeiro (Q_1) e o terceiro quartil (Q_3), permitindo identificar valores que estão significativamente distantes da maioria dos dados observados. A aplicação do IQR no tratamento de dados envolve as seguintes etapas:

- Cálculo dos quartis: primeiramente, são calculados o primeiro (Q_1) e o terceiro quartil (Q_3) da distribuição dos dados de consumo;
- Determinação dos limites: com os quartis calculados, estabelecem-se os limites superior e inferior. O limite inferior é dado por $Q_1 - 1{,}5 * IQR$ e o limite superior por $Q_3 + 1{,}5 * IQR$. Valores fora desses limites são considerados *outliers*;
- Tratamento dos *outliers*: identificados os *outliers*, estes podem ser removidos ou substituídos por valores mais representativos.

Ao aplicar o método IQR, é possível realizar uma análise mais precisa, eliminando o ruído causado por valores atípicos e garantindo que as conclusões sejam mais representativas da realidade.

2.3.2 Amostragem Representativa

Ademais, em estudos que possuem grandes bancos de dados, a quantidade massiva de informações pode introduzir desafios significativos em termos de armazenamento, processamento e até mesmo análise. Além disso, dados faltantes ou erros de medição podem comprometer a qualidade da investigação, levando a análises imprecisas. Nessa perspectiva, é comum e recomendado que seja selecionada uma amostra representativa do banco de dados para contornar esses obstáculos.

A amostragem, portanto, é uma técnica estatística que permite a análise de um subconjunto de dados para inferir conclusões sobre a população total. Quando o banco de dados é volumoso, uma amostragem bem feita facilita o manuseio dos dados e assegura resultados mais robustos e generalizáveis. Isto posto, é de interesse que a amostra retirada seja significativa para refletir de maneira fidedigna as características da área de estudo como um todo com um nível de confiança para a média pré-estabelecido.

Nesse contexto, para a retirada da amostra é preciso de algumas informações:

- Tamanho da população total: número total de grupos que se tem o banco de dados;
- Margem de erro: probabilidade de desvio da amostra em relação à população total;
- Nível de confiança: índice de certeza de que todo o grupo obteria a mesma resposta em um determinado intervalo.

$$Tamanho\ da\ amostra = \frac{\frac{z^2 * p * (1-p)}{E^2}}{1 + (\frac{z^2 * p * (1-p)}{E^2 * N})} \tag{2.1}$$

Onde,

- $z =$ Escore z;
- $p =$ Desvio padrão;
- $E =$ Margem de erro;
- $N =$ População total.

O escore z traduz o nível de confiança desejada em um valor numérico que representa quantos desvios padrão determinada proporção tem em relação à média. Esses valores de escore são mostrados na Tabela 2.3.

Tabela 2.3 - Escore z.

Nível de confiança desejado	Escore z
99%	2,58
95%	1,96
90%	1,65
85%	1,44
80%	1,28

2.4 Variações de Consumo

As flutuações no consumo são resultado de diversas variáveis externas, como clima, hábitos da população e temperatura. Assim, ao abordar o serviço de abastecimento de água, é crucial priorizar a consideração das variações diárias e horárias no consumo. A variação diária refere-se às discrepâncias observadas ao longo dos dias do ano, com aumentos durante o verão e reduções no inverno. Por sua vez, a variação horária lida com as flutuações ao longo das horas do dia, podendo ser diretamente influenciada pelo estilo de vida da população, entre outros fatores.

O coeficiente de variação diária (K_1) pode ser expresso tanto pela Figura 2.2, que apresenta a variação da demanda de água no período de um ano, quanto pela seguinte fórmula:

$$K_1 = \frac{maior\ vazão\ diária\ no\ ano\ (L/hab.dia)}{vazão\ média\ diária\ no\ ano\ (L/hab.dia)} \tag{2.2}$$

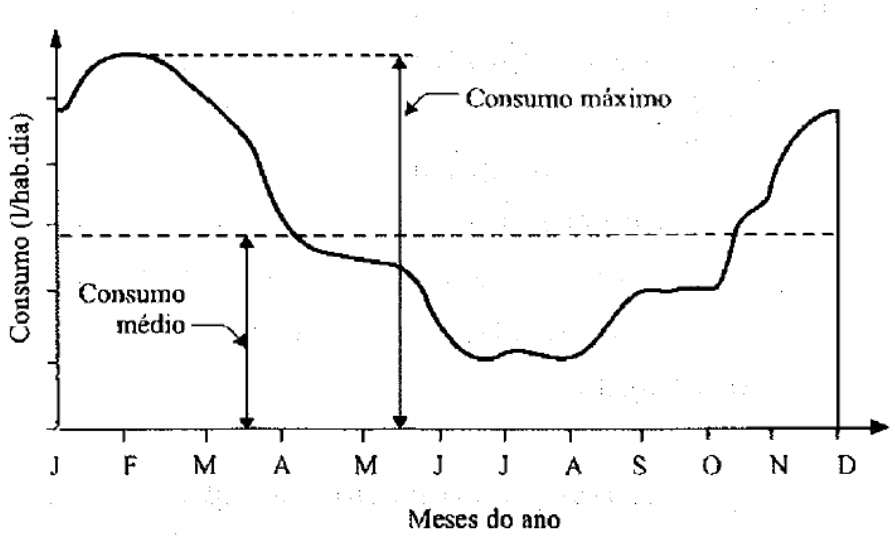


Figura 2.2 - Variações de consumo anual.

Fonte: (Tsutyia, 2004).

Ao determinar o coeficiente da hora de maior consumo (K_2), utiliza-se a seguinte fórmula:

$$K_2 = \frac{maior\ vazão\ horária\ no\ dia\ (L/s)}{vazão\ média\ do\ dia\ (L/s)} \quad (2.3)$$

Essa relação também pode ser visualizada na Figura 2.3.

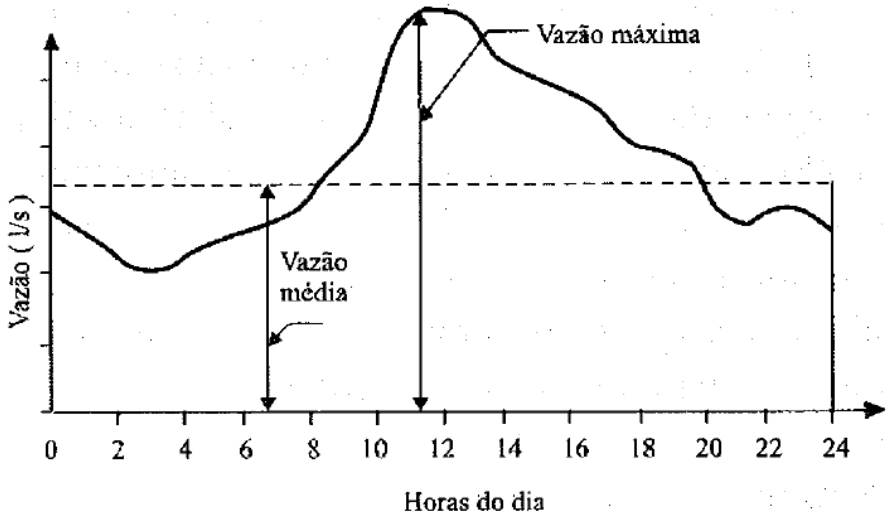


Figura 2.3 - Variações do consumo diário.

Fonte: (Tsutyia, 2004).

Assim, a relação entre a maior vazão diária (K_1) e o consumo médio diário no mesmo período, e a maior vazão horária (K_2) e o vazão média horária do mesmo dia

é demonstrada nas Tabela 2.4 e Tabela 2.5, respectivamente. Essas informações são derivadas de medições reais ou são recomendações pertinentes para o desenvolvimento de projetos.

Tabela 2.4 - Coeficiente do dia de maior consumo (K_1).

Autor/Entidade	**Local**	**Ano**	**Coeficiente K_1**	**Condições de obtenção do valor**
DAE	São Paulo Capital	1960	1,5	Recomendação para projeto
FESB	São Paulo Interior	1971	1,25	Recomendação para projeto
Azevedo Netto	Brasil	1973	1,1 - 1,5	Recomendação para projeto
Yassuda e Nogami	Brasil	1976	1,2 - 2,0	Recomendação para projeto
CETESB	Valinhos e Iracemápolis	1978	1,25 - 1,42	Medições em sistemas operando há vários anos
NB-587-ABNT	Brasil	1977	1,2	Recomendação para projeto
Orsini	Brasil	1996	1,2	Recomendação para projeto
Azevedo Netto *et al.*	Brasil	1998	1,1 - 1,4	Recomendação para projeto
Tsutiya	RMSP Setor Lapa	1989	1,08 - 3,8	Medições em sistema operando há vários anos
Saporta *et al.*	Barcelona Espanha	1993	1,1 - 1,25	Medições em sistema operando há vários anos
Walski *et al.*	EUA (*)	2001	1,2 - 3,0	Recomendação para projeto

Autor/Entidade	Local	Ano	Coeficiente K_1	Condições de obtenção do valor
Hammer	EUA (*)	1996	1,2 - 4,0	Medições em sistemas norte-americanos
AEP	Canada (*)	1996	1,5 - 2,5	Recomendação para projeto

(*) Nesses sistemas não há reservatórios domiciliares.

Fonte: (Tsutyia, 2004).

Tabela 2.5 - Coeficiente da hora de maior consumo (K_2).

Autor/Entidade	Local	Ano	Coeficiente K_1	Condições de obtenção do valor
Azevedo Netto	Brasil	1973	1,5	Recomendação para projeto
Yassuda e Nogami	Brasil	1976	1,5 - 3,0	Recomendação para projeto
CETESB	Valinhos e Iracemápolis	1978	2,08 - 2,35	Medições em sistemas operando há vários anos
PNB-587-ABNT	Brasil	1977	1,5	Recomendação para projeto
Orsini	Brasil	1996	1,5	Recomendação para projeto
Azevedo Netto *et al.*	Brasil	1998	1,5 - 2,3	Recomendação para projeto
Tsutiya	RMSP Setor Lapa	1989	1,5 - 4,3	Medições em sistema operando há vários anos
Saporta *et al.*	Barcelona Espanha	1993	1,3 - 1,4	Medições em sistema operando há vários anos

Autor/Entidade	Local	Ano	Coeficiente K_1	Condições de obtenção do valor
Walski *et al.*	EUA (*)	2001	3,0 - 6,0	Recomendação para projeto
Hammer	EUA (*)	1996	1,5 - 10,0	Medições em sistemas norte-americanos
AEP	Canada (*)	1996	3,0 - 3,5	Recomendação para projeto

(*) Nesses sistemas não há reservatórios domiciliares

Fonte: (Tsutyia, 2004).

Na revisão bibliográfica conduzida por Diao *et al.* (2010), examina-se a influência dos fatores de pico de vazão de água e suas características. Nesse contexto, evidencia-se que os valores dos coeficientes de consumo de água mantêm uma relação direta com o intervalo de tempo considerado. Em outras palavras, o fator de pico aumenta à medida que o intervalo de observação diminui, conforme destacado por Johnson (1999). Adicionalmente, destaca-se uma correlação significativa e inversamente proporcional entre o número de habitantes e a variação de consumo (Johnson *apud* Guidi, 2017).

Nesse contexto, diante da influência da densidade populacional na área de estudo sobre os coeficientes de consumo, há diversas abordagens para o cálculo destes, incluindo métodos variados e até mesmo a utilização de equações empíricas. Dentre esses métodos, em que P é a população, destacam-se inicialmente os coeficientes de variação de consumo diário (K_1):

- As relações alemãs propostas pela Associação Técnica e Científica Alemã para Gás e Água (DVGW) (citadas por Guidi, 2017), expressas por:

$$K_1 = -0{,}159 \ln P + 3{,}5488 \tag{2.4}$$

$$K_1 = 3{,}9 . P^{-0{,}0752} \tag{2.5}$$

- A fórmula empírica de Tricarico *et al.* (2007), derivada da análise estatística do consumo de água no sul da Itália, utilizando modelos log-normal e

Gumbel por meio da abordagem determinística que estabelece uma relação com o número de habitantes:

$$K_1 = 11.P^{-0,2} \quad (2.6)$$

Para calcular o coeficiente de consumo máximo horário, Guidi (2017) consolida os valores de K1 e K2 em um único coeficiente, representado por K, que resulta da multiplicação desses dois coeficientes para a análise do pico de vazão. Em seu embasamento teórico, ele faz referência às seguintes formulações para a determinação de K:

- Zhang (citado por Guidi, 2017) propõe o cálculo dos coeficientes de variação de consumo usando pulsos retangulares de Poisson, com fórmulas variando de acordo com a porcentagem de consumo interno, isto é, levando em consideração os diferentes usos de água entre usos ao ar livre (uso externo) e usos residenciais (internos), conforme apresentado a seguir:
    - Para 100% de consumo interno:

$$K = 2,5 + \frac{2,18}{\sqrt{P/_{1000}}} \quad (2.7)$$

    - Para 90% de consumo interno:

$$K = 3,02 + \frac{2,28}{\sqrt{P/_{1000}}} \quad (2.8)$$

    - Para 60% de consumo interno:

$$K = 4,17 + \frac{2,46}{\sqrt{P/_{1000}}} \quad (2.9)$$

- O modelo da AWWA, foi desenvolvido especificamente para pequenas comunidades e bairros individuais, ou seja, para populações de 650 a 1.675 habitantes, considerando o uso interno de 90%, em que "q" representa a demanda média per capita anual por 1000 pessoas em L/min.

$$K = \left(\frac{1095,31}{q}\right) * P^{-0,4} \quad (2.10)$$

- A DVGW também estabeleceu um método para o coeficiente K:

$$K = -0,75.\ln P + 11.679 \quad (2.11)$$

$$K = 18{,}1.\,\mathrm{P}^{-0{,}1682} \tag{2.12}$$

- A SABESP (citada por Guidi, 2017) recomenda que, para valores de vazão média inferiores a 751 L/s, adote-se K1= 1,2 e K2=1,5. No entanto, para valores superiores a esse limite, sugere-se a seguinte fórmula, que estabelece uma relação com a soma das vazões médias de uso residencial, comercial e público em litros por segundo (Q_m):

$$K = 1{,}2 + \frac{17{,}485}{{Q_m}^{0{,}5090}} \tag{2.13}$$

2.5 Modelagem de Consumo

No que tange o planejamento, operação e gerenciamento de sistemas de distribuição de água, um dos maiores pilares é a satisfação do consumidor final em aspectos como quantidade, qualidade e pressão. Todavia, a natureza estocástica da demanda de água em relação ao clima, a região geográfica e aos aspectos socioeconômicos de seus consumidores, influencia no seu padrão de consumo efetivo. Nesse aspecto, diversos autores citam e estudam a importância imprescindível da utilização de modelos de previsão de demanda de água a curto, médio e longo prazo.

De acordo com Donkor, Mazzuchi, *et al.* (2012), existem diversas variáveis medidas em diferentes períodos que influenciam na previsão da demanda de água urbana. Desse modo, diferentes métodos são apresentados na literatura, com diferentes objetivos, determinantes, horizonte de projeto e periodicidade.

2.5.1 Modelos Aplicados à Dados de Macromedição

Segundo Tsutyia (2004), a demanda de água pode ser determinada através de três métodos: leitura dos hidrômetros (micromedição), leitura do macromedidor instalado na saída do reservatório (macromedição) e quando não existe medição.

Neste contexto, a obtenção do cálculo do consumo médio efetivo per capita requer a disponibilidade de dados de consumo específicos para cada tipo de economia (residencial, comercial, industrial e público), juntamente com o número total de unidades em cada categoria. Essas informações são geralmente coletadas em

cada leitura dos hidrômetros realizada pela concessionária. O cálculo então pode ser feito através da seguinte equação:

$$q_e = \frac{V_c}{NE.ND.NH/L} \tag{2.14}$$

Tal que,

- $q_e =$ Consumo efetivo *per capita* de água, em L/s;
- $V_c =$ Volume consumido medido pelos hidrômetros, em L/s;
- $NE =$ Número médio de economias;
- $ND =$ Número de dias da medição pelos hidrômetros;
- $NH/L =$ Número de habitantes por ligação.

É crucial salientar a inclusão no cálculo do valor das perdas de água ao longo do percurso da tubulação, tendo em mente que essas não são totalmente estanques (Tsutyia, 2004), como expresso na equação a seguir:

$$q = \frac{q_e}{1 - I} \tag{2.15}$$

Onde,

- $q =$ Consumo *per capita* de água, em L/s;
- $I =$ Índice de perdas.

Ao realizar a leitura do macromedidor, se os dados de volume consumidos no período são registrados por hora, torna-se viável estimar os coeficientes de variação de vazão e construir um gráfico do tipo tempo-vazão. A Figura 2.4 ilustra um exemplo desse tipo de gráfico, exibindo a variação do perfil de consumo para o Setor Itaim Paulista.

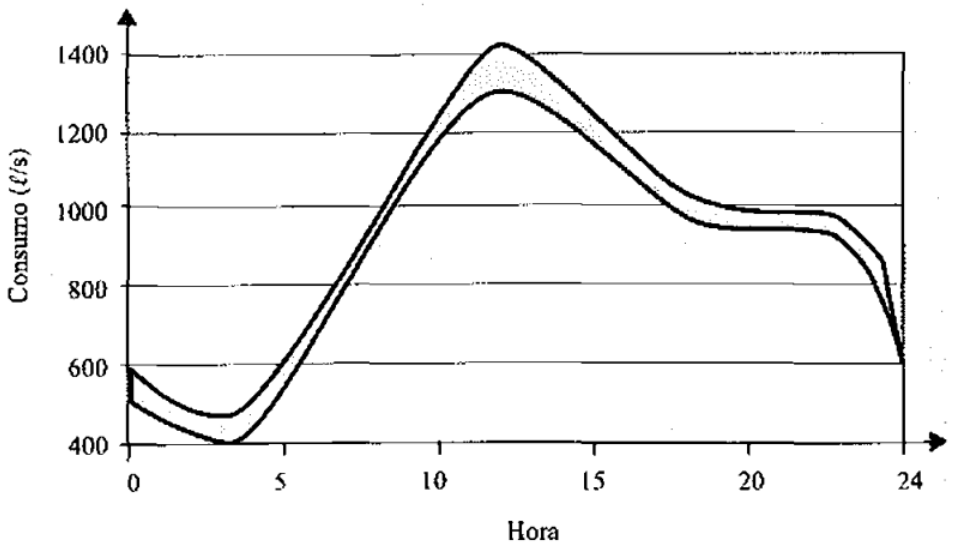


Figura 2.4 - Variação do perfil de consumo em função do dia da semana para o setor Itaim Paulista.

Fonte: (Tsutyia, 2004).

Entretanto, quando não há medição é preciso adotar valores de consumo médio *per capita* por tipo de estabelecimento e coeficientes de variação de vazão encontrados em diferentes literaturas (Tsutyia, 2004).

Nesse ínterim, Tsutyia (2004) sugere que o dimensionamento da rede de distribuição de água deve ser conduzido considerando as condições mais críticas, ou seja, a demanda máxima. Isso visa assegurar o desempenho adequado do sistema em todas as horas do dia e todos os dias do ano. A expressão para o cálculo da vazão do reservatório até a rede é fornecida por:

$$Q = \frac{K_1 . K_2 . P . q}{86.400} + Q_{esp} \tag{2.16}$$

Tal que,

- Q = Vazão, em L/s;
- K_1 = Coeficiente do dia de maior consumo;
- K_2 = Coeficiente da hora de maior consumo;
- P = População da área abastecida, em hab.;
- q = Consumo per capita, em L/hab.dia;
- Q_{esp} = Vazão específica, em L/s.

Ademais, Guidi (2016) empregou uma análise de regressão linear multivariada, uma técnica estatística que desenvolve modelos com base nas relações entre

variáveis explicativas de um processo. A abordagem envolveu 14 variáveis, incluindo, mas não se limitando a temperatura, nebulosidade, dia da semana e vazão na hora anterior. Essas variáveis foram categorizadas em quatro grupos distintos, visando analisar a distribuição da vazão no sistema de abastecimento de água no Setor Ana Dorothéa (Zona Baixa) na cidade de Franca, São Paulo. Dessa forma, em um modelo de regressão múltipla, a variável dependente (Y) será determinada por mais de uma variável independente (X), podendo ser representada por:

$$Y_i = \alpha + \sum_{k=1}^{p} \beta_k X_{k_i} + e_i \tag{2.17}$$

- $\alpha =$ Valor esperado de Y quando todas as variáveis independentes forem nulas;
- $\beta_k =$ Variação esperada de Y dado um incremento unitário em X_k, mantendo todas as demais variáveis independentes constantes;
- $e_i =$ Erro não explicado.

2.5.2 Modelos Aplicados à Dados de Micro e Macromedição

Além disso, a análise apropriada e acurada da demanda de água a curto prazo pode ser realizada por meio de séries temporais univariadas (UTS), regressão de séries temporais (TSR), rede neural artificial (RNA) e métodos híbridos (Anele, Hamam, *et al.*, 2017). No entanto, este estudo foca exclusivamente em métodos que oferecem uma abordagem mais direta para a compreensão de séries temporais, em alinhamento com o objetivo principal discutido anteriormente. Por essa razão, os modelos de RNA e métodos híbridos, que demandam maior complexidade, poder computacional e quantidade de dados para treinamento, não serão abordados em profundidade.

Séries temporais univariadas (UTS) são conjuntos de dados que consistem em observações passadas, coletadas ao longo do tempo para uma única variável e que apresentam como propriedade a estacionariedade, isto é, sua média, variância e covariância não sofrem variações com o tempo. Nesse aspecto, os modelos preditivos para esse tipo de análise são baseados em observações do passado e nos seus erros associados. Os modelos UTS chave são suavização exponencial, autorregressão (AR), média móvel (MA), média móvel autorregressiva (ARMA) e média móvel autorregressiva integrada (ARIMA). Segundo Anele, Hamam, *et al.* (2017), mesmo

esses modelos sendo bastante utilizados para a previsão operacional à curto prazo, também podem não ser tão precisos sob o efeito de variáveis exógenas, como mudanças climáticas que afetam a demanda de água.

A suavização exponencial é um modelo linear que utiliza ponderações exponenciais para atribuir diferentes pesos a observações passadas, proporcionando uma ênfase decrescente aos dados mais antigos. Essa abordagem é particularmente útil quando a ênfase nas observações recentes é mais relevante para a previsão futura do que as observações mais antigas, entretanto é importante reconhecer as limitações inerentes ao método de suavização exponencial. Conforme destacado por Anele, Hamam, *et al.* (2017), os modelos de processos estocásticos AR, MA, ARMA e ARIMA são utilizados a partir do momento que a suavização exponencial deixa de ser adequada devido ao perfil complexo das séries temporais em estudo (apud Donkor, Mazzuchi, *et al.*, 2012, p. 154).

A autorregressão (AR), por sua vez, assume que o valor atual de uma variável depende linearmente de seus valores passados e de um termo de erro estocástico (ruído branco). Um modelo autorregressivo de ordem *p* AR(p) é baseado na combinação linear de *p* valores passados para realizar uma previsão, tal como segue na equação.

$$\Upsilon_t = \mu + \sum_{k=1}^{p} \phi_k \, \Upsilon_{t-k} + \epsilon_t \tag{2.18}$$

Em que,

- Υ_t = Valor de demanda de água observado no tempo t;
- μ = Média;
- p = Ordem da média móvel;
- ϕ = Parâmetro autorregressivo;
- k = Número de períodos históricos;
- ϵ_t = Erro de previsão no tempo t.

A média móvel (MA), modelo também linear, utiliza *q* erros dos valores passados e sua combinação linear em relação à média para previsão de demanda futuras. O modelo de média móvel MA(q) pode ser descrito como:

$$\Upsilon_t = \mu + \epsilon_t + \sum_{k=1}^{q} \theta_k \, \epsilon_{t-k} \tag{2.19}$$

Onde,

- q = Ordem da média móvel;
- θ = Parâmetro de média móvel.

A combinação dos modelos de médias móveis e os modelos autorregressivos levam ao desenvolvimento do modelo de média móvel autorregressivo (ARMA), portanto, este modelo, para a previsão de demandas de água da série temporal, leva em consideração os valores observados em adendo aos erros passados. O processo do modelo ARMA segue a seguinte equação:

$$\Upsilon_t = \mu + \sum_{k=1}^{p} \phi_k \Upsilon_{t-k} + \epsilon_t + \sum_{k=1}^{q} \theta_k \epsilon_{t-k} \quad (2.20)$$

Assim como o modelo de média móvel autorregressiva (ARMA), o modelo de média móvel integrada autorregressiva (ARIMA) é concebido pela combinação dos modelos autorregressivo (AR) e de média móvel (MA). No entanto, o ARIMA vai além ao considerar explicitamente a falta de estacionariedade nos dados da série temporal. Portanto, para avaliar se a série é estacionária, sem considerar variações pontuais ou saltos, é possível empregar o teste de Dickey-Fuller, que é um teste de raiz unitária. A análise se fundamenta na hipótese nula (H_0) de que o parâmetro δ é igual a zero, indicando que a série é não estacionária e há uma raiz unitária. Em contrapartida, a hipótese alternativa (H_1) sugere que a série é estacionária, uma vez que não há uma raiz unitária. O teste pode ser expresso pela seguinte equação:

$$\Delta Y_t = \delta Y_{t-1} + \epsilon_t \quad (2.21)$$

Tal que,

- ΔY_t = Diferença entre os valores sucessivos da série temporal;
- δ = Parâmetro a ser testado;
- Y_{t-1} = Valor anterior.

No entanto, quando se busca identificar degraus ou quebras estruturais, decorrentes de mudanças significativas no comportamento da série temporal, como alterações abruptas na média ou tendência, é recomendável utilizar testes mais específicos, como o teste de Chow, Zivot-Andrews, Bai-Perron e Quandt-Andrews. Esses métodos são mais adequados para capturar e analisar tais rupturas no padrão dos dados.

Nesse ínterim, caso a série seja não estacionária e como os modelos discutido anteriormente necessitam da estacionaridade para serem efetivamente aplicados, é necessário a conversão da série para uma série temporal estacionária por meio da diferenciação de ordem *d*, componente “I” do método ARIMA, assim como segue:

$$I = \Upsilon'_t = \Upsilon_t - \Upsilon_{t-1} \tag{2.22}$$

Em que,

- Υ'_t = Componentes das diferenças tomadas da série.

O modelo ARIMA, portanto, trata-se da combinação do modelo autorregressivo (AR), da diferenciação (I) e do modelo de médias móveis, formando assim uma abordagem abrangente para a modelagem de séries temporais.

$$\phi_p(B)(1-B)^d\Upsilon_t = \theta_q(B)\epsilon_t \tag{2.23}$$

Tal que,

- B = Operador da defasagem (lag);
- $\phi_p(B)$ = Polinômio de auto-regressão de ordem p;
- $\theta_q(B)$ = Polinômio de média móvel de ordem q;
- d = Número de diferenciações.

Nos cenários em que se identifica a presença de sazonalidade na série temporal, é apropriado empregar o modelo de média móvel integrada autorregressiva sazonal (SARIMA). A notação do modelo SARIMA(p,d,q)(P,D,Q)s envolve a especificação da ordem dos modelos AR(p), I(d) e MA(q) como parâmetros para uma função ARIMA, adicionando ainda os parâmetros AR(P), I(D), MA(Q) e *s* para a componente sazonal, onde "s" representa o número de observações por sazonalidade.

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D\Upsilon_t = \Theta_Q(B^s)\theta_q(B)\epsilon_t \tag{2.24}$$

Onde,

- B = Operador da defasagem (lag);
- $\Phi_P(B^s)$ = Polinômio de auto-regressão sazonal de ordem P;
  $\Theta_Q(B^s)$ = Polinômio de média móvel sazonal de ordem Q;
- D = Número de diferenciações sazonais.

Vale ressaltar que os métodos de Série Temporal Univariada (UTS) mesmo podendo trabalhar com perfis complexos, não levam em consideração os efeitos de variáveis externas como o clima e o valor das tarifas de água. (Anele, Hamam, *et al.*, 2017). Portanto, a principal crítica a estes métodos reside na omissão de considerar os impactos das alterações demográficas, climáticas, econômicas e tecnológicas, bem como das variáveis e estratégias de gestão da demanda de água ao avaliar o futuro consumo de água (Donkor, Mazzuchi, *et al.*, 2012).

2.5.3 Avaliação e Seleção de Modelos

Para assegurar uma escolha apropriada dos modelos de previsão de demanda de água, é crucial considerar os erros de previsão que desempenham um papel significativo na avaliação, tal que acurácia do modelo em relação aos dados observados é determinada pela comparação entre ambos. No entanto, de acordo com Donkot, Mazzuchi, *et al.* (2012), para a melhor seleção do modelo, é necessário seguir as seguintes etapas:

1. Divisão do conjunto de dados em um período de treino e um período de teste;
2. Utilização do período de estimativa para modelar a demanda;
3. Avaliação da acurácia dos modelos, comparando as previsões com os valores observados tanto para o período de treino quanto para o período de teste;
4. Seleção do melhor modelo com base no desempenho, conforme medido por qualquer uma das medidas de erro especificadas a seguir.

A seleção do modelo deve considerar tanto o desempenho no ajuste à série de treino quanto a precisão nas previsões para a série de teste. Para isso, as métricas de avaliação são divididas de maneira a refletir esses dois aspectos, garantindo uma análise equilibrada entre a capacidade do modelo de aprender com os dados históricos e sua habilidade de generalizar para dados futuros.

2.5.3.1 Avaliação do Ajuste

Assim como proposto por Costa (2022) e Anele, Hamam, *et al.* (2017), pode-se utilizar o Critério de Informação Akaike (AIC), originalmente proposto por Akaike

(1974) para avaliar a acurácia de cada modelo e seu ajuste a curto prazo. O método de AIC pode ser representado:

$$AIC = N.\log\left(\frac{RSS}{N}\right) + 2k \tag{2.25}$$

Em que,

- $RSS =$ Número de observações;
- $RSS =$ Resíduo estimado do modelo ajustado;
- $k =$ Número de parâmetros estimados para o modelo.

De acordo com Costa (2022), o critério em questão está associado à teoria da informação e tem a capacidade de quantificar a perda de informação de um modelo. Ele oferece uma abordagem simples para a seleção do melhor modelo, sendo aquele que apresenta o menor índice, uma vez que o método equilibra a qualidade do ajuste com a complexidade do modelo, penalizando modelos com muitos parâmetros para evitar o *overffiting*. Isso implica que a escolha do melhor modelo está diretamente relacionada à sua capacidade de se aproximar dos dados, resultando na menor perda de conhecimento.

Adicionalmente, o Teste de Ljung-Box deve ser aplicado para verificar a independência dos resíduos. Nesse teste, a hipótese nula (H_0) postula que os resíduos do modelo não apresentam autocorrelação. Um valor-p superior a 0,05 indica que a hipótese nula não pode ser rejeitada, sugerindo que os resíduos são puramente aleatórios e não correlacionados, o que reforça a adequação do modelo ao capturar corretamente as estruturas subjacentes dos dados.

O cálculo do teste é realizado por meio da análise da autocorrelação de uma série de resíduos em diferentes defasagens ao mesmo tempo, em que o valor estatístico Q, que segue uma distribuição qui-quadrado com m graus de liberdade, é definido como:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{r}_k^2}{n-k} \tag{2.26}$$

Em que,

- $n =$ Quantidade de dados;
- $m =$ Número de defasagens a serem testadas;
- $\hat{r}_k^2 =$ Autocorrelação dos resíduos na defasagem k.

2.5.3.2 Avaliação do Ajuste e Previsão

O Erro Médio Absoluto (MAE) representa a diferença média entre os valores modelados e os valores observados para uma determinada base de dados. Sua fórmula é definida por:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| \tag{2.27}$$

Onde,

- $Y_t =$ Valor observado;
- $\hat{Y}_t =$ Valor modelado.

Por sua vez, o Erro Médio Quadrático (MSE) mensura a média do quadrado da diferença entre os valores observados e os valores reais. Sua expressão é dada por:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2 \tag{2.28}$$

Além disso, também pode-se empregar a Raiz do Erro Médio Quadrático (RMSE), calculada pela seguinte equação:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2} \tag{2.29}$$

Portanto, para a comparação da performance dos modelos de previsão, o modelo com o menor valor em uma medida de erro escolhida é considerado mais acurado.

Ademais, vale também citar o Coeficiente de Eficiência de Nash-Sutcliffe (NSE), métrica estatística muito utilizada para validar modelos de previsão em hidrologia, principalmente, mas também em outras ciências ambientais. Dessa forma o NSE mede a habilidade do modelo em prever valores em comparação com a média dos valores observados. Portanto o coeficiente é dado por:

$$NSE = 1 - \frac{\sum_{t=1}^{n}\left(Y_t - \hat{Y}_n\right)^2}{\sum_{t=1}^{n}(Y_t - \bar{Y}_n)^2} \tag{2.30}$$

Onde,

- $\bar{Y} =$ Média dos valores observados.

Os valores do NSE fornecem uma métrica clara do desempenho do modelo: um NSE igual a 1 indica um ajuste perfeito, enquanto um valor maior que zero mostra

que o modelo performa melhor do que a média dos valores observados. Um NSE igual a zero significa que o modelo tem o mesmo desempenho que a média dos valores observados, e qualquer valor negativo aponta para um desempenho inferior ao da média.

Todavia, para avaliar o desempenho de modelos de regressão linear, além das métricas já citadas, utilizam-se também o coeficiente de determinação (R^2) e o erro padrão residual.

O coeficiente de determinação (R^2) é uma métrica que indica a proporção da variabilidade dos dados que é explicada pelo modelo de regressão. Seu valor varia de 0 a 1, sendo que valores mais próximos de 1 indicam um ajuste melhor, ou seja, uma maior capacidade do modelo de prever os valores observados. A fórmula para o cálculo do R^2 é dada por:

$$R^2 = 1 - \frac{\sum\left(Y_t - \hat{Y}_t\right)^2}{\sum((Y_t - \bar{Y}_t)^2} \tag{2.31}$$

O erro padrão residual (RSE), por sua vez, mede a precisão das previsões do modelo, refletindo a média das distâncias verticais entre os pontos observados e a linha de regressão ajustada, tal que resíduos menores indicam melhores previsões. O erro padrão residual segue a formulação:

$$RSE = \sqrt{\frac{\sum\left(Y_t - \hat{Y}_t\right)^2}{n - p}} \tag{2.32}$$

Em que,

- $p =$ Número de variáveis explicativas no modelo.

3 METODOLOGIA

3.1 Caracterização da Área de Estudo

A Região Administrativa do Lago Norte, RA nº XVII do Distrito Federal, foi formalmente estabelecida pela Lei nº 641, de 10/01/1994, e teve suas diretrizes regulamentadas pelo Decreto nº 15.516/94. Em seu início, a RA estava subordinada à RA I - Plano Piloto, o que a coloca sob as restrições e parâmetros de ocupação do Plano Piloto, devido a sua proximidade e vinculação urbanística. A Região Administrativa apresenta uma área total de 7.549,17 hectares e seu abastecimento de água é realizado pelos reservatórios do Torto, de Santa Maria, do Lago Paranoá e do Ribeirão Bananal. A área do Lago Norte pode ser visualizada na Figura 3.1.

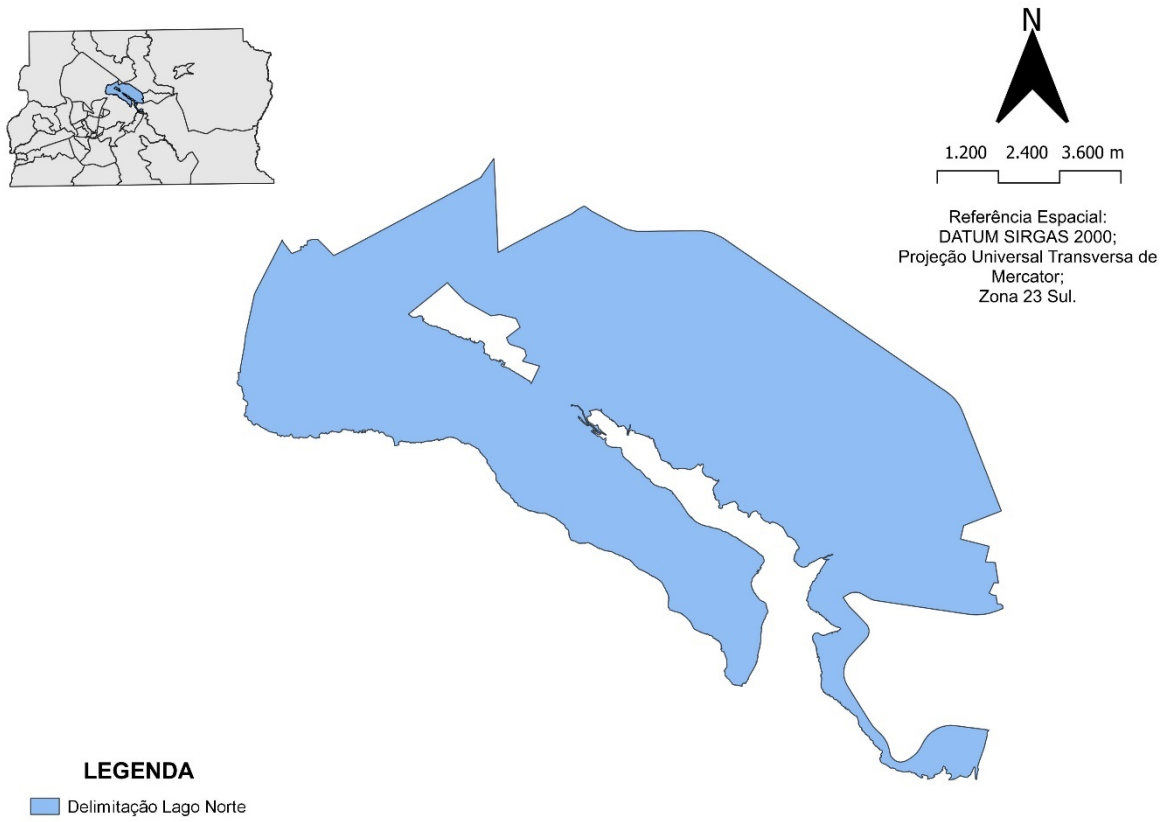


Figura 3.1 - Localização da Região Administrativa do Lago Norte.

A Região Administrativa do Lago Norte abrange a Península Norte, isto é o Setor Habitacional Individual Norte (SHIN) e o Setor de Mansões do Lago Norte (SMLN), o Setor Habitacional Taquari, o Centro de Atividades e os Núcleos Rurais Remanescentes, como o Núcleo Rural do Córrego do Torto, Núcleo Rural do Córrego Jerivá e Núcleo Rural do Córrego do Urubu. Nessa configuração, a RA oferece uma variedade de infraestrutura, incluindo hospitais, parques ambientais, shoppings

centers, espaços turísticos, escolas, postos de saúde e uma delegacia (CODEPLAN, 2021). No entanto, é importante ressaltar que a área é primariamente residencial, como pode ser visualizado na Figura 3.2.

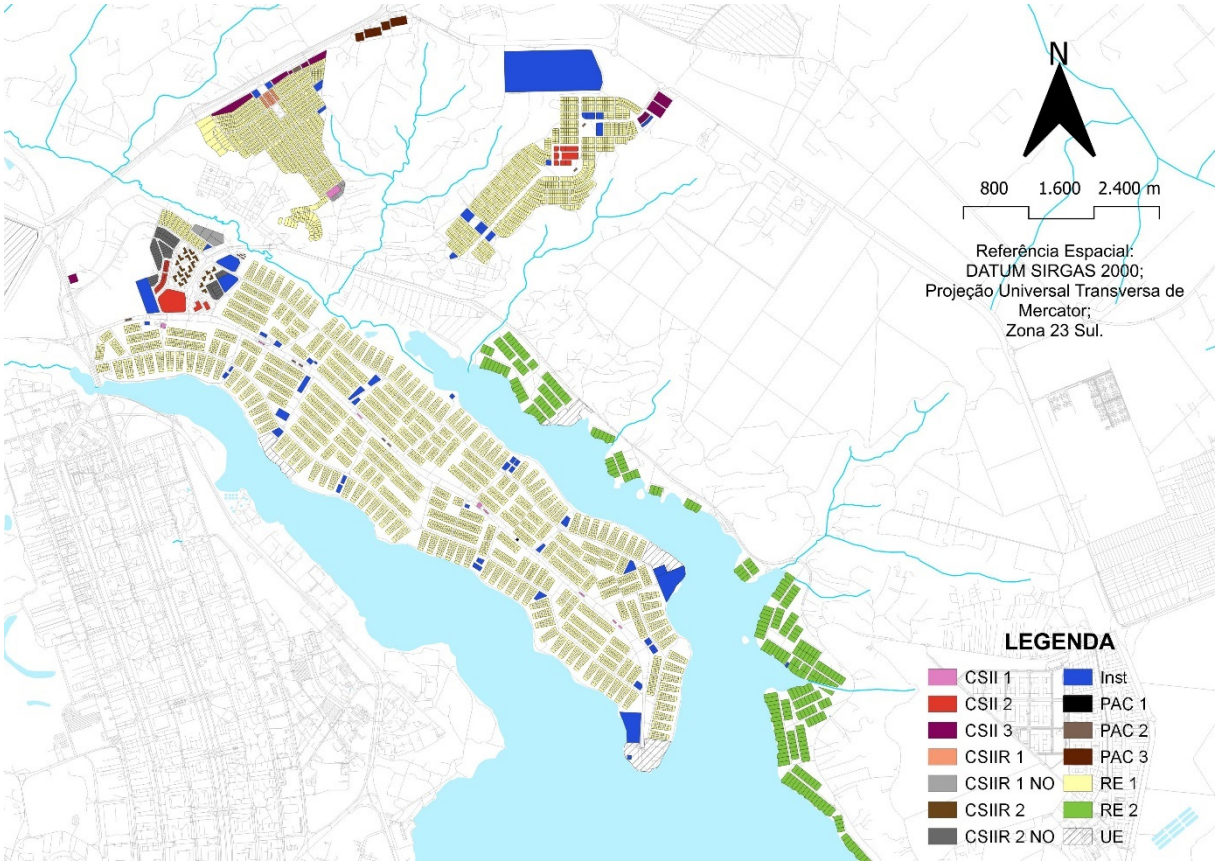


Figura 3.2 - Uso e ocupação da RA Lago Norte.

O PDAD 2021 revela que a população urbana do Lago Norte atingiu 37.539 habitantes, com uma renda domiciliar estimada em R$ 15.867,10, resultando em uma média per capita de R$ 6.489,00. A pesquisa também estima 13.987 unidades de domicílio particular ocupadas, com uma média de 2,7 moradores por domicílio. A condição permanente foi observada em 97,7% dos domicílios, com a maioria situada em casas fora de condomínio. A Figura 3.3 apresenta a distribuição dos domicílios ocupados de acordo com o tipo.

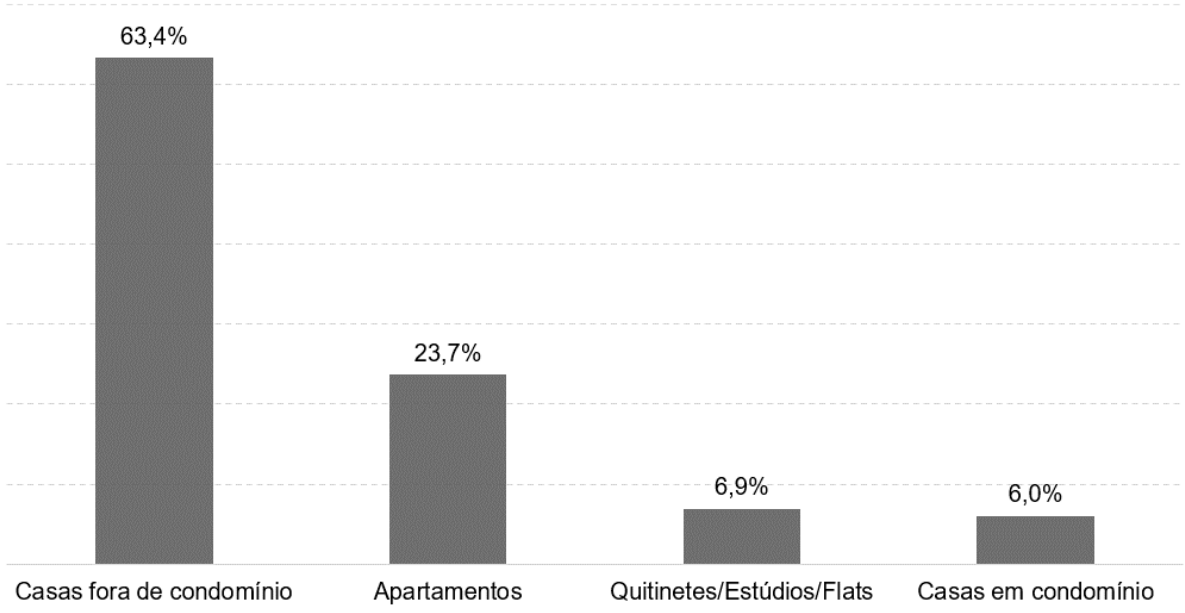


Figura 3.3 - Distribuição dos domicílios ocupados segundo o tipo.

Fonte: (CODEPLAN, 2021).

No que diz respeito ao abastecimento de água, destaca-se que 94,9% dos domicílios contavam com acesso à rede geral da Companhia de Saneamento Ambiental do Distrito Federal (CAESB), enquanto 3% utilizavam poço/cisterna e 1,7% tinham como fonte um poço artesiano. Adicionalmente, 96,4% afirmaram possuir uma caixa d'água (CODEPLAN, 2021). A Tabela 3.1 apresenta a contagem de domicílios com abastecimento de água nas residências enquanto a Figura 3.4 exibe o respectivo gráfico.

Tabela 3.1 - Abastecimento de água no domicílio.

Variável	Total		%	
	Sim	Não	Sim	Não
Rede geral (CAESB)	13.188	709	94,9	5,1
Poço/Cisterna	489	13.376	3,5	96,5
Poço artesiano	236	13.630	1,7	98,3
Captação de água da chuva	(***)	13.850	(***)	99,9
Gambiarra/gato	(***)	13.850	(***)	99,8
Caixa d'água	13.140	489	96,4	3,6

Fonte: (CODEPLAN, 2021).

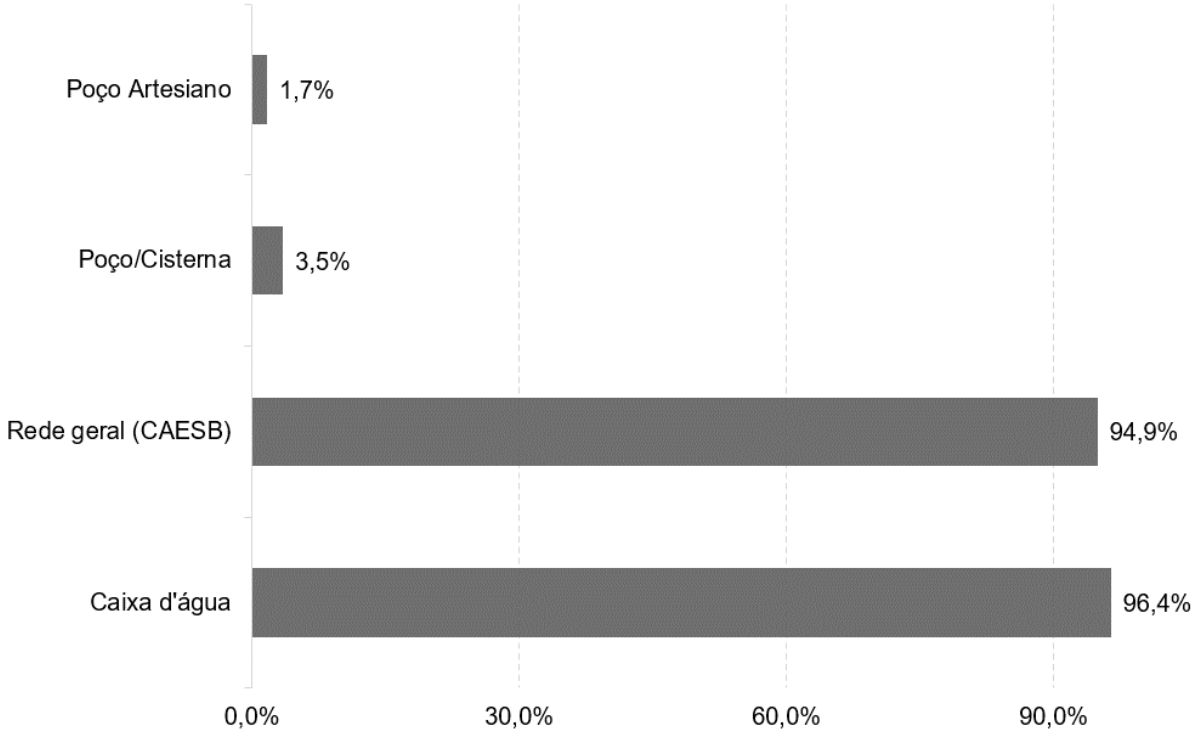


Figura 3.4 - Abastecimento de água no domicílio.

Fonte: (CODEPLAN, 2021).

A Região Administrativa do Lago Norte se destacou como pioneira na implementação de um sistema avançado de telemetria. Desde as Quadras Intermediárias (QI) até as Quadras do Lago (QL) 1 a 16, foram instalados estrategicamente dispositivos eletrônicos nos hidrômetros de cada residência. Esses

dispositivos realizaram leituras precisas do consumo de água pelo menos quatro vezes ao dia, de 11 de junho de 2019 a 3 de março de 2023, enviando automaticamente os dados, de forma segura, por meio da rede de telefonia celular para os servidores da Companhia de Saneamento Ambiental do Distrito Federal (CAESB). Nesse ínterim, para os propósitos deste trabalho, a área de estudo abrange exclusivamente a Península Norte, onde os sistemas de medição inteligente já foram plenamente implementados, conforme demonstrado na Figura 3.5.

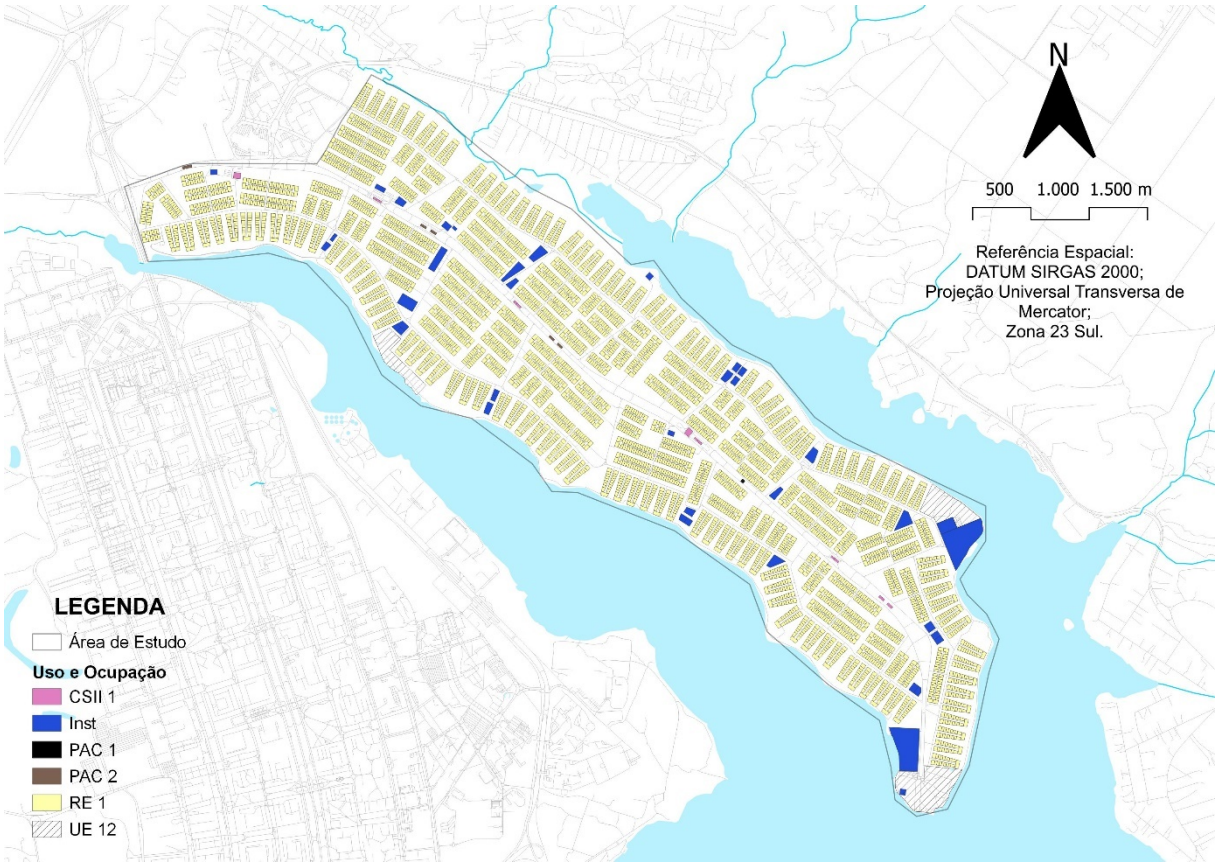


Figura 3.5 - Área de estudo.

A tecnologia implementada não apenas permite uma coleta de dados frequente, mas também viabiliza a análise detalhada dos consumos diários. Este avanço é crucial para a identificação e controle de padrões de uso, proporcionando à CAESB a capacidade de adotar medidas de operação e prevenção eficientes. Além disso, também é possível detectar de forma ágil os vazamentos nas tubulações domiciliares da Região Administrativa (Caesb, 2019).

O sistema de telemetria, impulsionado por um investimento aproximado de 1,7 milhão de reais do Banco Interamericano de Desenvolvimento, representa um marco significativo no aprimoramento de medição de consumo de água. Esse financiamento

foi direcionado para beneficiar tanto a Companhia de Saneamento Ambiental do Distrito Federal (CAESB) quanto os consumidores finais, impactando aproximadamente 6 mil clientes por meio de uma inovadora tecnologia baseada na Internet das Coisas (IoT). Essa é a primeira vez que a IoT é implementada em larga escala no setor de saneamento básico brasileiro, representando um avanço notável.

O sistema de distribuição do Lago Norte é composto por um reservatório e a infraestrutura de distribuição propriamente dita, incluindo tubulações e diversos acessórios. O controle preciso do volume de água fornecido à região é realizado por meio de dois macromedidores de vazão estrategicamente instalados a jusante do reservatório de distribuição. A configuração hidráulica detalhada da área de estudo pode ser visualizada na Figura 3.6. Essa representação gráfica oferece uma visão clara da disposição e interconexão dos componentes do sistema, contribuindo para uma compreensão abrangente da infraestrutura hídrica do Lago Norte.

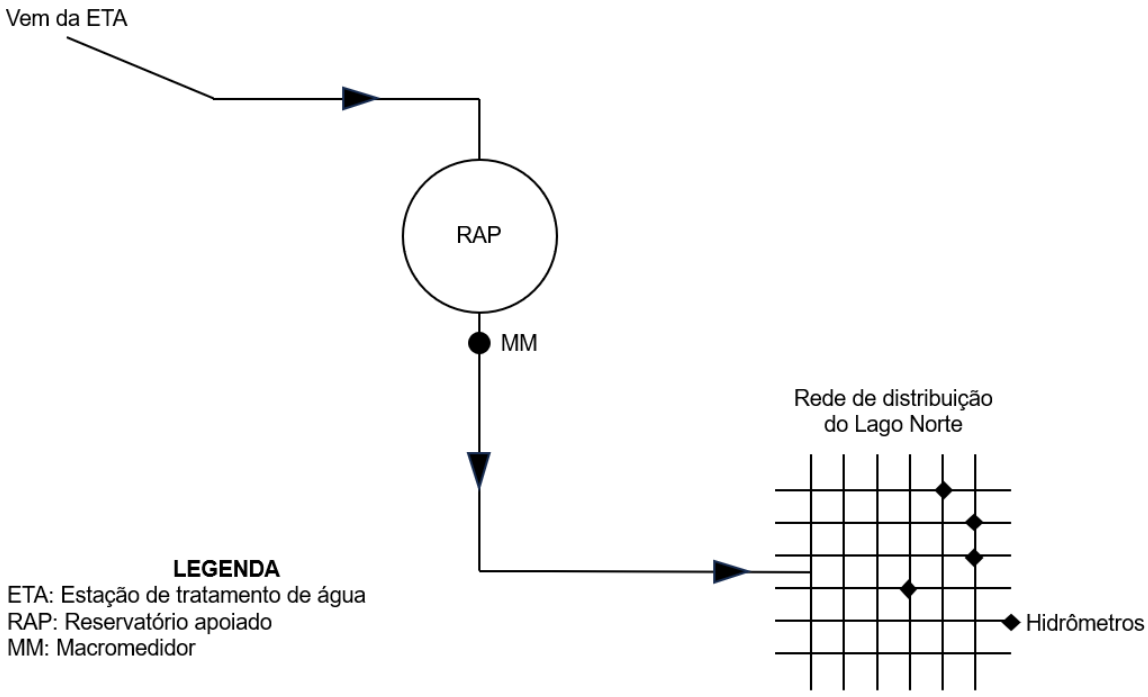


Figura 3.6 - Configuração hidráulica do sistema de abastecimento de água do Lago Norte.

No Lago Norte, atualmente, são registradas 6.692 ligações prediais, das quais 5.608 estão equipadas com hidrômetros dotados de sistema de telemetria, representando, portanto, R$ 303,13 por ponto, baseado no investimento total para a região. Os hidrômetros empregados nessa região pertencem à classe de exatidão 1 e são do tipo ultrassônico, apresentando uma faixa de vazão máxima variável entre 3,0

m³/h e 65 m³/h, sendo a maioria deles configurada para uma vazão de 3,0 m³/h. Todos os hidrômetros estão pré-equipados para a instalação de telemetria. A Figura 3.7 ilustra o modelo de hidrômetro utilizado na área de estudo.



Figura 3.7 - Exemplo de hidrômetro da área de estudo.

Os hidrômetros implantados na Península Norte são do modelo *DIEHL Metering* da empresa *Hydrus.* Segundo o catálogo da empresa (HYDRUS, 2024) esses hidrômetros são de alta precisão, feitos de latão sem chumbo para instalações residenciais, comerciais e industriais e contam com as seguintes características:

- Precisão extrema em fluxos de baixa vazão e estabilidade de medição a longo prazo;
- Detecção de vazamento integrada;
- Montável em qualquer posição de instalação;
- Corpo de liga de cobre sem chumbo;
- Armazenamento de dados horários por 42 dias;
- Protocolo *Diehl Extended Encoder* que inclui temperatura, alarmes e mensagens de erro, entre outros.

Os dados técnicos gerais do hidrômetro utilizado podem ser melhor observadas na Tabela 3.2.

Tabela 3.2 - Dados técnicos gerais do hidrômetro DIEHL Metering da Hydrus.

Dado	Informação
Temperatura da água potável	1ºC a 50ºC
Temperatura do ambiente de operação	-25ºC a 55ºC
Temperatura de armazenamento ambiente	-25ºC a 70ºC (32ºC no máximo para 1 hora)
Pressão máxima de operação	2,068e+6 Pa
Bateria	2 baterias 3.6 VDC de lítio
Tempo de vida da bateria	Mais de 20 anos
Faixa de vazão operacional	0,023 a 7,27 m³/h
Faixa de baixa vazão	0,009 a 0,023 m³/h
Vazão inicial	0,004 m³/h
Precisão da faixa operacional	±1,5
Precisão da faixa de baixa vazão	±5
Perda de pressão	13,79 kPa para 3,41 m³/h

Fonte: Adaptado (HYDRUS, 2024).

Outrossim, as dimensões do medidor, conforme o catálogo da *Hydrus* são expostas na Figura 3.8 e na Tabela 3.3.

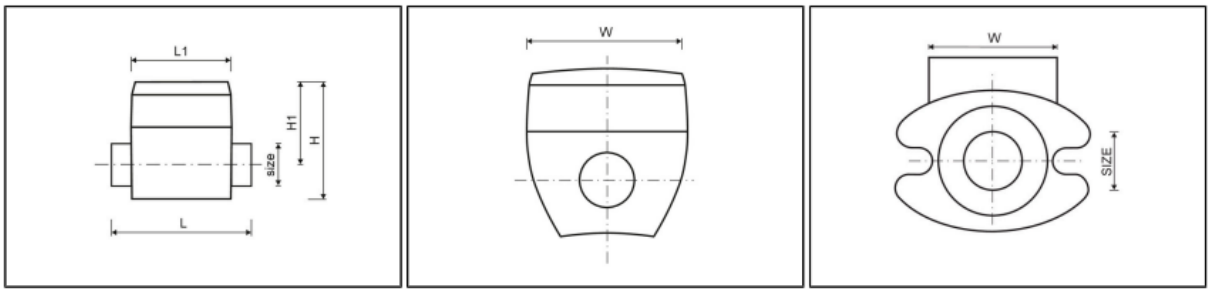


Figura 3.8 - Dimensões do hidrômetro DIEHL Metering da Hydrus.

Fonte: (HYDRUS, 2024).

Tabela 3.3 - Dimensões do hidrômetro DIEHL Metering da Hydrus.

Dimensão	Valor
Comprimento de instalação	228,6 mm
Comprimento do registro (L1)	88,9 mm
Largura do registro (W)	93,98 mm
Altura total (H)	101,6 mm
Altura do centro do tubo até o topo (H1)	68,58 mm

Fonte: Adaptado (HYDRUS, 2024).

O sistema de medição inteligente implementado nos hidrômetros da área de estudo representa uma inovação significativa, permitindo o monitoramento em tempo real das vazões de consumo ao longo do dia. Essa abordagem não apenas otimiza a operação da rede de distribuição de água, mas também promove a conscientização e a responsabilização dos consumidores, ao envolvê-los de maneira mais ativa em relação ao seu consumo hídrico.

No contexto da Península Norte, a infraestrutura de medição compreende 5.608 hidrômetros equipados com rádios de transmissão de leitura e 5 *gateways* responsáveis por receber os dados provenientes desses dispositivos. A rede de comunicação adotada para o sistema de telemetria é composta por tecnologia *LoRaWan* para a comunicação interna e 3G/4G para a comunicação externa. O sistema supervisor utilizado é o *WMNet*, desenvolvido pela empresa contratada para a instalação e gestão dos sistemas de telemetria. Essa configuração oferece uma estrutura robusta para a coleta, transmissão e supervisão eficiente dos dados de vazão, promovendo uma gestão mais inteligente e eficaz dos recursos hídricos na região.

3.2 Modelagem do Consumo de Água

Com base na fundamentação teórica e nos dados de medição fornecidos pela Companhia de Saneamento Ambiental do Distrito Federal (CAESB), foram desenvolvidos modelos para a análise do consumo de água, adotando a mesma metodologia proposta por Costa (2022).

Todavia, inicialmente, antes de aplicar os modelos, foi necessário realizar o tratamento dos dados brutos para garantir a precisão e eficiência desejadas para o

estudo. Este processo envolveu a limpeza de dados, visando remoção de inconsistências e erros que poderiam comprometer a integridade da análise e a correção de outros, quando possível. Este processamento foi fundamental para seguir com confiança na análise e aplicação dos modelos.

Nesse contexto, três modelos foram empregados: o modelo ARIMA, SARIMA e a regressão linear múltipla multivariada. O escopo da modelagem visa avaliar o padrão de consumo no distrito de medição do Lago Norte, compreendendo as variações diárias e horárias, além de possibilitar a utilização dos dados de telemetria para desenvolver uma ferramenta que apoie o dimensionamento de redes de água, seguindo uma abordagem semelhante à do presente estudo de caso. É relevante destacar que a simulação de todos os modelos foi executada por meio do software *RStudio*.

É importante destacar que, para o estudo, foi realizada uma análise abrangente de uma amostra representativa da Península Norte. O objetivo foi extrapolar os dados de consumo para toda a região, permitindo a aplicação subsequente dos modelos de previsão com base em uma amostra que reflete as características gerais da área.

3.2.1 Modelo ARIMA

Com base nos dados de consumo, organizados para cada hidrômetro da amostra representativa da área de estudo e posteriormente agrupados em valores médios diários, criou-se a série temporal de vazão para análise e modelagem da região de estudo como um todo. O método de determinação dos parâmetros (p, d, q) do modelo ARIMA (Equação 2.23) é apresentado detalhadamente a seguir:

1. Divisão da base de dados em conjunto de treino (80% do total) e conjunto de teste (20% do total);
2. Aplicação do teste de *Dickey-Fuller* para avaliação da estacionariedade da série temporal;
3. Estabelecimento dos parâmetros do modelo pelo pacote auto.arima do *RStudio*;
4. Implementação da simulação do Modelo ARIMA;
5. Avaliação do desempenho do ajuste;
6. Análise dos resultados do AIC;
7. Análise da correlação de resíduos pelo teste de Ljung-Box;

8. Realização, por meio do *RStudio*, do cálculo dos valores de MAE (Erro Absoluto Médio), RMSE (Erro Quadrático Médio Quadrático) e NSE (Nash-Sutcliffe) para avaliar tanto o ajuste quanto a previsão.

3.2.2 Modelo SARIMA

Assim como no modelo ARIMA, o modelo SARIMA também se fundamentará na série temporal média geral da região de estudo. Entretanto, neste modelo, será incorporada a sazonalidade semanal. O processo detalhado para a determinação dos parâmetros (p, d, q) (P, D, Q) s do método SARIMA (Equação 2.24) é descrito a seguir:

1. Divisão da base de dados em conjunto de treino (80% do total) e conjunto de teste (20% do total);
2. Aplicação do teste de *Dickey-Fuller* para avaliação da estacionariedade da série temporal;
3. Estabelecimento dos parâmetros do modelo pelo pacote auto.arima do *RStudio*;
4. Implementação da simulação do Modelo SARIMA;
5. Avaliação do desempenho do ajuste;
6. Análise dos resultados do AIC;
7. Análise da correlação de resíduos pelo teste de Ljung-Box;
8. Realização, por meio do *RStudio*, do cálculo dos valores de MAE (Erro Absoluto Médio), RMSE (Erro Quadrático Médio Quadrático) e NSE (Nash-Sutcliffe) para tanto o ajuste quanto a previsão.

3.2.3 Modelo de Regressão Linear Multivariada

Dado que a Regressão Linear Multivariada incorpora variáveis explicativas além dos dados temporais das vazões de consumo, serão empregados procedimentos distintos em comparação com os modelos anteriores, visando identificar as relações entre as variáveis propostas (Costa, 2022). Desta forma, para este modelo, serão seguidos os seguintes passos:

1. Por meio dos dados de vazão médio diários, serão organizadas em um banco de dados outras variáveis, como: intervalo do dia para quando o dia foi computado, a vazão de consumo média, a vazão de consumo no intervalo anterior e as vazões para o mesmo dia nas semanas anteriores;

2. Análise da correlação entre as variáveis para definir a significância de cada uma para o modelo de regressão linear multivariada. Para isso será calculada a matriz de correlação com os valores do coeficiente de Pearson;
3. Processamento das variáveis significativas;
4. Parametrização estatística da equação que descreve a relação entre as variáveis explicativas e a vazão de consumo;
5. Avaliação do desempenho do modelo a partir dos valores de AIC, R^2, RSE e NSE.

4 RESULTADOS

4.1 Tratamento de Dados

Inicialmente, foram recebidos dados de vazão da Companhia de Saneamento Ambiental de Brasília (Caesb), abrangendo o período de 11 de junho de 2019 a 1 de março de 2023, distribuídos em arquivos .csv organizados por ano. No entanto, observou-se que os dados de 2019, 2020 e 2021 não eram representativos. Isso se deve ao fato de que a empresa responsável pela manutenção e monitoramento dos medidores inteligentes na região de estudo iniciou a troca de todos os aparelhos de telemetria da região, que se mostraram defeituosos, apenas em agosto de 2021. Como resultado, os dados desses anos, anteriores à troca dos equipamentos, foram descartados por não serem adequados para o estudo.

Adicionalmente, o contrato de licitação com a empresa responsável pelos medidores inteligentes tinha vigência até março de 2023. Por conta da proximidade do término do contrato, a empresa não realizou as manutenções necessárias, o que resultou em anomalias nos dados de 2023. Assim, esses dados também foram desconsiderados. Portanto, para o presente estudo, foram considerados apenas os dados do ano de 2022 disponibilizados pela Caesb, totalizando, ainda sim, 44.948.697 de observações para realizar a análise.

Com isso, deu-se início ao tratamento dos dados. Os dados fornecidos continham oito colunas: identificador da leitura (IDE_TAG), data da leitura (DTA_TAG), valor da leitura (VLR_TAG_LEI), nome da leitura (NOM_TAG), identificador do equipamento (IDE_EQU), data de ativação do ponto (DTA_ATI_PTO), inscrição do imóvel (INS_IMO) e auxiliar de inscrição do imóvel (INS_IMO_AUX). Contudo, algumas dessas colunas não eram relevantes para o estudo e foram excluídas, sendo elas as colunas referentes ao nome da leitura, à data de ativação do ponto e ao auxiliar de inscrição do imóvel, uma vez que este último apresentava valores idênticos aos da inscrição do imóvel.

Para facilitar a organização, as colunas remanescentes foram renomeadas e as unidades dos dados foram reatribuídas, visto que muitas delas haviam perdido sua unidade original ao serem importadas do arquivo .csv. A partir disso, o identificador foi definido como um valor numérico, a inscrição do imóvel foi configurada como um valor inteiro, a data foi ajustada para o formato POSIX no padrão Ano-Mês-Dia Hora-Minuto-Segundo, e a leitura foi definida como um valor numérico. Vale ressaltar que o formato

POSIX é um padrão amplamente utilizado para representar data e hora de maneira consistente e legível dentro do *RStudio*.

Em seguida, foi realizada a limpeza inicial dos dados, removendo duplicatas, valores ausentes (NA) e leituras negativas do *dataframe*. Além disso, para garantir que apenas os dados transmitidos pelos medidores substituídos em 2021 fossem mantidos, os dados foram filtrados com base no identificador do equipamento, considerando que os medidores atualizados começavam com "70B3". Como resultado dessas etapas de limpeza e filtragem, o conjunto de dados foi reduzido para 43.366.619 registros, 1.582.078 observações a menos.

Prosseguiu-se com o ordenamento dos dados por imóvel e data de leitura, seguido pela retirada de uma amostra com um nível de confiança de 95% e uma margem de erro de 0,05, selecionando uma amostra de 360 imóveis dos 5678 presentes no *dataframe*, abrangendo, assim, um total de 6.668.776 dados. Por fim, foi realizada a criação de intervalos regulares de 4 horas, uma vez que a maioria das leituras seguia esse padrão com o objetivo de facilitar a análise.

Em seguida, foram identificados e corrigidos os *spikes* nos dados, por meio do método do intervalo interquartil (IQR), tal que os valores identificados como anômalos foram substituídos por NA. Posteriormente, esses campos *Not Available* foram preenchidos utilizando progressão aritmética entre o primeiro e o último valor numérico adjacente a cada NA. Nos casos em que essa substituição não foi possível, pela falta de números adjacentes a montante e/ou a jusante do valor não disponível, o campo ficou sem preenchimento.

A análise do intervalo interquartil (IQR) dos dados revelou a presença de uma anomalia significativa no *dataframe*. O IQR, que reflete a dispersão dos valores entre o primeiro e o terceiro quartil apontou valores extremamente elevados. Esse resultado indica uma variabilidade excessiva dentro do conjunto de dados, sugerindo que existem ainda *spikes* significativos influenciando a distribuição, tal que estes não foram identificados pelo método IQR.

Portanto, em seguida, o volume de consumo em m³ foi calculado subtraindo a leitura anterior da leitura atual. Isto posto, foram eliminados quaisquer valores de volume de consumo negativos, já que esses indicam um decréscimo entre leituras consecutivas, que, na prática, não ocorre em medições volumétricas de consumo de hidrômetros, demonstrando falha nos aparelhos ou na transmissão de informação.

Assim, a quantidade de dados baixou, chegando a 6.564.746. Os valores de volume de consumo calculados podem ser observados na Tabela 4.1.

Tabela 4.1 - Valores de volume de consumo entre leituras consecutivas na amostra.

Valores	Volume de consumo (m³)
Min.	0,000
1º Qu.	0,039
Mediana	0,555
Média	5,984
3º Qu.	1,987
Max.	3201,800

Concluiu-se que ainda havia *spikes* impactando a análise, evidenciados por valores de consumo de até 3.201,8 m³ entre leituras consecutivas. Considerando que a maioria das leituras ocorre a cada 4 horas, esse valor é claramente injustificável. Para abordar esse problema, os dados foram analisados em busca de possíveis erros, e verificou-se que, na maioria dos casos, esses picos resultaram de *gaps* entre leituras dentro de cada imóvel, onde o consumo de vários dias foi acumulado em uma única leitura subsequente.

Nesse ínterim, dentro da amostra representativa, calculou-se a média do volume de consumo para cada imóvel. Assim, utilizando os valores de referência fornecidos por Tsutyia (2004) para o Setor Lapa da RMSP, conforme detalhado nas Tabela 2.4 e Tabela 2.5, foram identificados e substituídos os valores de volume consumo anômalos. Para isso, estabeleceu-se um limite para a identificação de *outliers*, calculado como o produto $K_1 * K_2 * Média\ por\ imóvel$, onde K_1 e K_2 são coeficientes determinados, conforme discutido no item 2.4, resultando em um valor de $3{,}8 * 4{,}3 * Média\ por\ imóvel$. Os valores que ultrapassaram esse limite foram substituídos por NA.

Com os *outliers* identificados, procedeu-se ao cálculo da média de volume de consumo para cada intervalo de 4 horas em cada imóvel. Essas médias foram utilizadas para preencher os valores ausentes (NA), garantindo a continuidade e a

consistência dos dados. Esse processo finalizou o tratamento dos dados fornecidos, resultando nos valores apresentados na Tabela 4.2.

Tabela 4.2 - Valores de volume de consumo entre leituras consecutivas na amostra após tratamento.

Valores	Volume de consumo (m³)
Min.	0,0000
1º Qu.	0,0010
Mediana	0,0040
Média	0,0359
3º Qu.	0,0180
Max.	20,6000

Embora ainda tenha sido possível identificar a presença de valores abruptamente altos, esses tiveram um impacto mínimo na média geral diária da região de estudo. Como o foco deste estudo é a modelagem da média geral dos dados da amostra, a influência desses valores foi insignificante, representando apenas cerca de 0,4% dos 6.564.746 registros amostrais. Com isso, a média de consumo foi calculada para cada dia, e os dados tratados foram exportados em um arquivo .csv.

Tabela 4.3 - Valores de volume de consumo médio diários na amostra.

Valores	Volume de consumo (m³)
Min.	0,01200
1º Qu.	0,02700
Mediana	0,03700
Média	0,03673
3º Qu.	0,04400
Max.	0,08900

4.2 Modelagem de Consumo

4.2.1 Série Temporal

Com os dados tratados, iniciou-se a modelagem do consumo. Os dados foram importados para o *RStudio*, onde foi necessário realizar uma transformação inicial para calcular o consumo diário em L/h. Considerando que o volume de consumo estava em m³/dia, a conversão foi feita dividindo o volume por 24 horas, correspondente a um dia, e multiplicando o resultado por 1.000 para converter de m³ para litros.

Após a conversão dos dados, eles foram transformados em uma série temporal utilizando o pacote *ts* do R. Esse pacote cria objetos de séries temporais no R, que podem ser vetores ou matrizes e que herdam a classe *ts*. Isso permite que esses objetos realizem operações aritméticas e alinhem os eixos temporais de forma adequada, além de incluir atributos adicionais que representam dados amostrados em unidades de tempo.

O argumento *frequency* do pacote *ts* especifica a frequência de amostragem da série temporal. Neste caso, foi utilizado o valor 7, uma vez que os dados são amostrados diariamente e a sazonalidade esperada da série é de uma semana. Dessa forma, foi possível plotar o autocorrelograma e o autocorrelograma parcial da série temporal, representados na Figura 4.1.

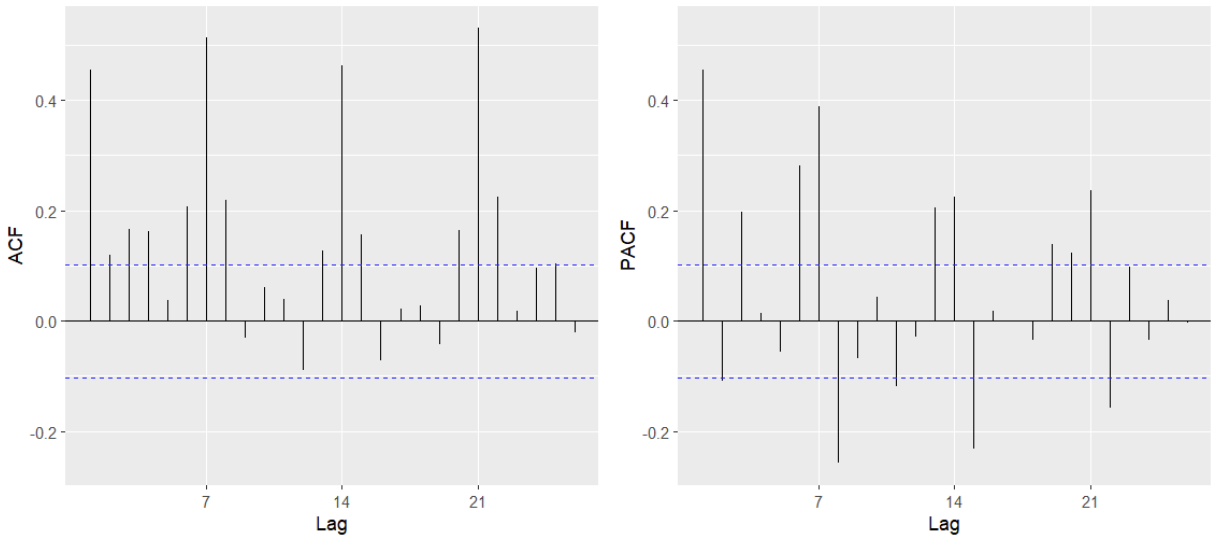


Figura 4.1 - Autocorrelograma e autocorrelograma parcial da série temporal.

A ACF mede o grau de correlação entre uma observação da série temporal e suas versões defasadas em diferentes lags, enquanto a PACF elimina a influência dos lags intermediários, permitindo avaliar a correlação direta entre uma observação e

uma defasagem específica, sem interferência de outras. Dessa forma, os gráficos resultantes são usados para sugerir o número de lags relevantes.

Além disso, para assegurar a integridade dos dados e garantir maior segurança na análise, foi utilizado o pacote *tsclean*, que tem a função de identificar e substituir *outliers* e valores faltantes em séries temporais. O pacote emprega a metodologia de outro pacote do R, denominado *supsmu*, para suavização de séries. Para substituir valores faltantes e outliers, é aplicada a interpolação linear, que ajusta os dados de forma contínua e precisa ao longo da série temporal. A limpeza realizada não identificou nenhum *outlier* novo.

Neste contexto, procedeu-se à verificação de estacionariedade da série temporal utilizando o pacote *adf.test*, que executa o Teste de *Dickey-Fuller*. Este teste avalia a hipótese nula de que uma série temporal univariada possui uma raiz unitária, o que indicaria que a série é não estacionária. O valor-p obtido foi de 0,01, que é inferior a 0,05, levando à rejeição da hipótese nula.

Dado que a série foi confirmada como estacionária com nível de confiança de 95%, o critério de diferenciação para os modelos ARIMA e SARIMA pode ser ajustado para zero, pois a diferenciação não é necessária. Assim, apesar da série ser estacionária, a modelagem prosseguiu com ARIMA e SARIMA para garantir a adequação completa do modelo, considerando que esses modelos incluem componentes adicionais que podem ser úteis para capturar padrões mais complexos, além da simples estacionariedade.

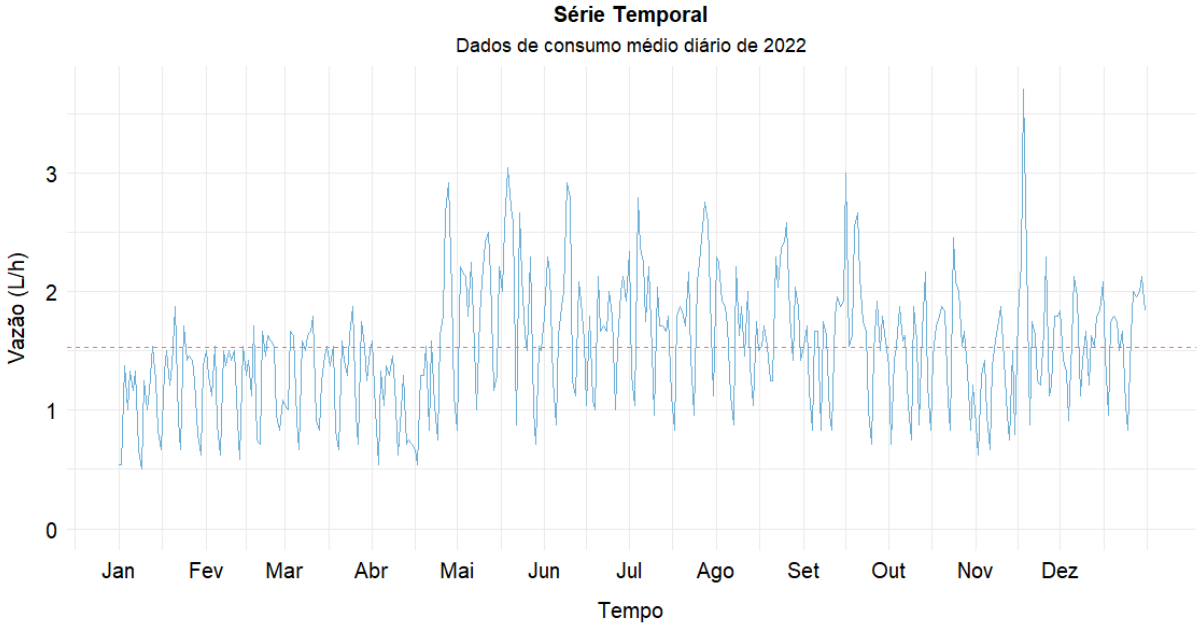


Figura 4.2 – Série temporal e vazão média dos dados observados.

Assim, pode-se observar que a série temporal ajustada da Figura 4.2 apresenta valores de consumo inferiores à média (representada pela linha vermelha tracejada) de janeiro a abril de 2022. Contudo a partir de maio, nota-se que os valores de consumo variam em torno da média, tanto para mais quanto para menos. Ao realizar uma análise pormenorizada dos motivos dessa discrepância, observou-se que a quantidade de dados se manteve constante ao longo dos meses, girando em torno de uma média de 548.896 dados mensais. Todavia, ao analisar os imóveis minuciosamente, foi perceptível que a partir de maio, alguns imóveis realmente apresentaram valores de volume de consumo elevados, portanto, elevando a vazão média concomitantemente.

No entanto, de maneira geral, os valores de consumo não se desviam significativamente da média de 1,530 L/h, mantendo-se dentro de uma variação constante ao longo do tempo, tanto positiva quanto negativa. observa-se que a variação máxima diária em novembro atingiu 3,708 L/h, o que representa um aumento de 2,42 vezes em relação à vazão média, valor esse que aponta uma variação duas vezes maior do que a fixada na literatura para projetos, isto é, K_1 = 1,2. Assim, ao comparar a variação máxima diária calculada com os valores observados e as formulações de K_1 discutidas no Item 2.4, chegou-se aos seguintes resultados:

Tabela 4.4 - Comparativo de K_1.

Dados	AWWA 1	AWWA 2	Tricarico
2,42	2,97	1,77	1,34

Ao realizar o comparativo entre as variações máximas diárias de vazão utilizando diferentes abordagens, os resultados obtidos mostram discrepâncias significativas entre os métodos. O valor observado de 2,42 está mais próximo do método AWWA 1, que apresentou uma variação de 2,97. Já o método AWWA 2 apresentou um valor consideravelmente mais baixo, enquanto o método de Tricarico, por sua vez, resultou em uma variação de 1,34. Comparativamente, o modelo de Tricarico apresenta uma estimativa mais conservadora, sendo menor que os métodos AWWA, mas ainda assim superior à média da variação observada. Essas diferenças sugerem que os métodos adotados podem influenciar diretamente a análise de capacidade e dimensionamento dos sistemas hidráulicos.

Nessa perspectiva, com a série temporal e a análise de variações de consumo diária finalizadas, partiu para a modelagem de consumo.

4.2.2 Modelo ARIMA

Para iniciar a modelagem, a série temporal foi dividida em duas partes: uma série de treino, que correspondeu a 80% dos dados, e uma série de teste, que representou os 20% restantes. Com essa divisão, utilizou-se o pacote *auto.arima* do *RStudio* para determinar os coeficientes do modelo, tal que ele seleciona automaticamente o melhor modelo ARIMA com base no menor valor do Critério de Informação de Akaike (AIC), que indica a menor perda de informação possível.

O *auto.arima* realiza uma busca entre diversas configurações de modelos, considerando as informações fornecidas. Ao definir o parâmetro *stepwise* = FALSE, a função foi configurada para explorar todas as possíveis combinações de modelos, mesmo que isso implique um processo mais demorado, garantindo assim a escolha do modelo mais adequado. Portanto, para o presente estudo, obteve-se um modelo ARIMA(5,1,0), em que os coeficientes resultantes foram:

Tabela 4.5 - Coeficientes modelo ARIMA.

Coeficiente	Valor	Erro padrão	Valor-p
ar1	-0,3926	0,0535	2,206e-13
ar2	-0,5541	0,0571	<2,2e-16
ar3	-0,3705	0,0620	2,356e-09
ar4	-0,2168	0,0571	0,0001446
ar5	-0,4070	0,0537	3,464e-14

Os valores p encontrados são todos significativamente baixos, menores que 0,05, indicando que os coeficientes são estatisticamente significativos e demonstrando que para o modelo ARIMA o valor previsto depende dos últimos 5 valores passados. No entanto, apesar de o teste de *Dickey-Fuller* indicar que a série temporal é estacionária, o pacote *auto.arima* sugeriu um modelo com ordem de integração 1. Essa recomendação pode refletir a necessidade de incluir uma diferenciação sazonal que não foi identificada na análise inicial da série. A inclusão dessa diferenciação pode ser crucial para capturar padrões sazonais que não são evidentes na análise de estacionariedade ou para evitar o *overfitting*, promovendo um ajuste mais robusto e geral do modelo.

Desse modo, foram feitos testes do ajuste do modelo em relação aos valores observados, utilizando as métricas de avaliação AIC, MAE, RMSE e NSE, assim como detalhado na Tabela 4.6.

Tabela 4.6 - Métricas de avaliação do ajuste ARIMA.

AIC	MAE	RMSE	NSE
370,51	0,3643	0,4462	0,3234

O critério de informação de Akaike (AIC) de 370,51 indica que o modelo ARIMA(5,1,0) consegue equilibrar adequadamente a complexidade do modelo com a adaptação aos dados. O AIC penaliza a complexidade do modelo e um valor menor sugere um ajuste mais preciso, porém é essencial comparar esse valor com outros modelos para uma avaliação abrangente. Por outro lado, o RMSE de 0,4462 representa a magnitude média dos erros do ajuste. Com esse valor relativamente significativo, podemos inferir que o modelo consegue capturar algumas das variações nos dados, porém não é capaz de eliminar completamente os erros.

O MAE de 0,3643 reflete a média dos erros absolutos entre os valores previstos e observados. Valores menores de MAE indicam menor desvio médio e reforçam a precisão dos dados modelados. Já o NSE de 0,3234 aponta para uma eficiência moderada para baixa do modelo que sugere que o modelo consegue capturar uma parte mínima da variabilidade nos dados ainda e necessita de melhorias notáveis na captura da dinâmica completa do consumo.

As métricas AIC, RMSE e MAE do modelo ARIMA(5,1,0) sugerem que o modelo se ajusta singelamente aos dados observados, tal que o valor do NSE de 0,3234 aponta para uma eficiência limitada, talvez devido à ausência de sazonalidades ou outros fatores não incluídos na modelagem. Conclui-se, que o ARIMA não captura todas as nuances da série temporal, portanto, é recomendável considerar métodos alternativos ou ajustes adicionais. Dessa forma, o modelo ajustado pode ser melhor observado na Figura 4.3.

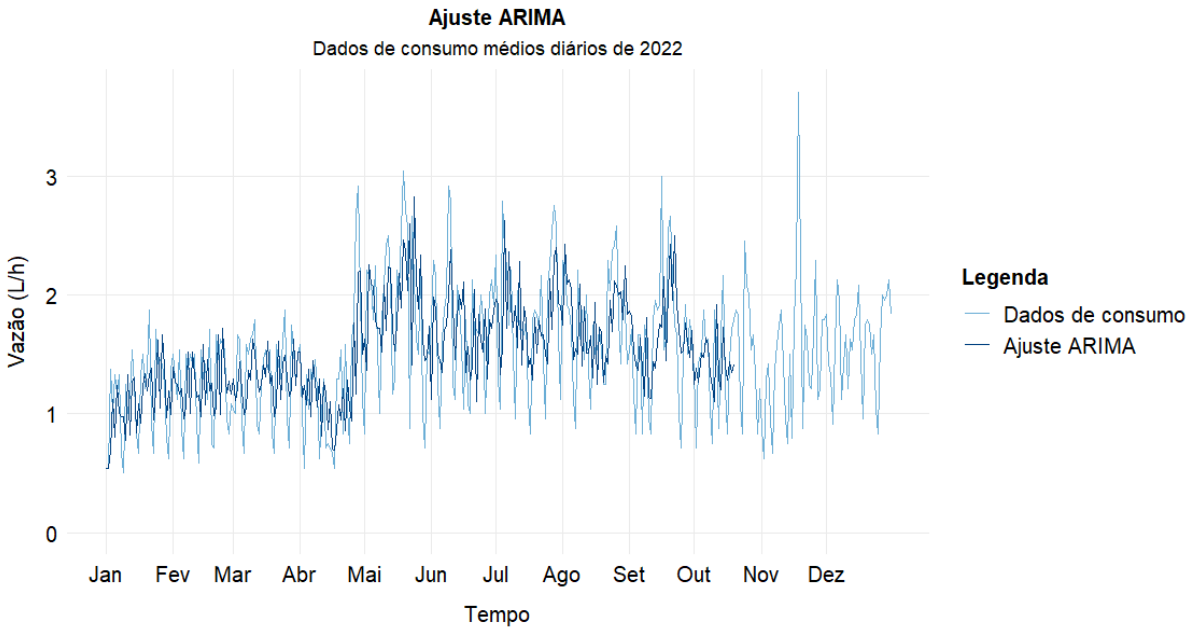


Figura 4.3 - Ajuste ARIMA.

Ao analisar os resíduos do modelo em relação aos dados, por meio do pacote *coeftest*, o teste de *Ljung-Box* apresentou um valor-p de 2,265e-14, que é significativamente menor que 0,05. Isso indica que há correlação entre os resíduos, sugerindo que eles não se comportam como um ruído branco. Portanto, a ausência de ruído branco nos resíduos sugere que o modelo pode não ter capturado todas as dinâmicas subjacentes da série temporal, uma vez que, em um modelo bem ajustado, os resíduos deveriam se comportar de tal forma, representando que restou apenas a aleatoriedade não explicada. Ademais, também foi realizada a verificação gráfica dos resíduos por meio do histograma e do gráfico de dispersão em relação a vazão, representados na Figura 4.4 e Figura 4.5.

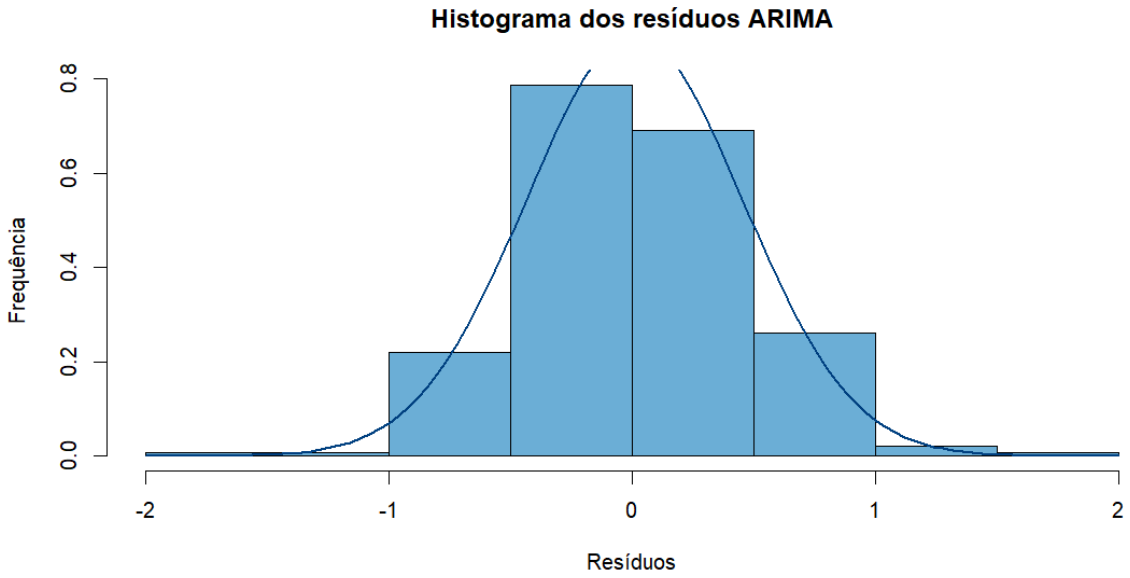


Figura 4.4 - Histograma dos resíduos ARIMA.

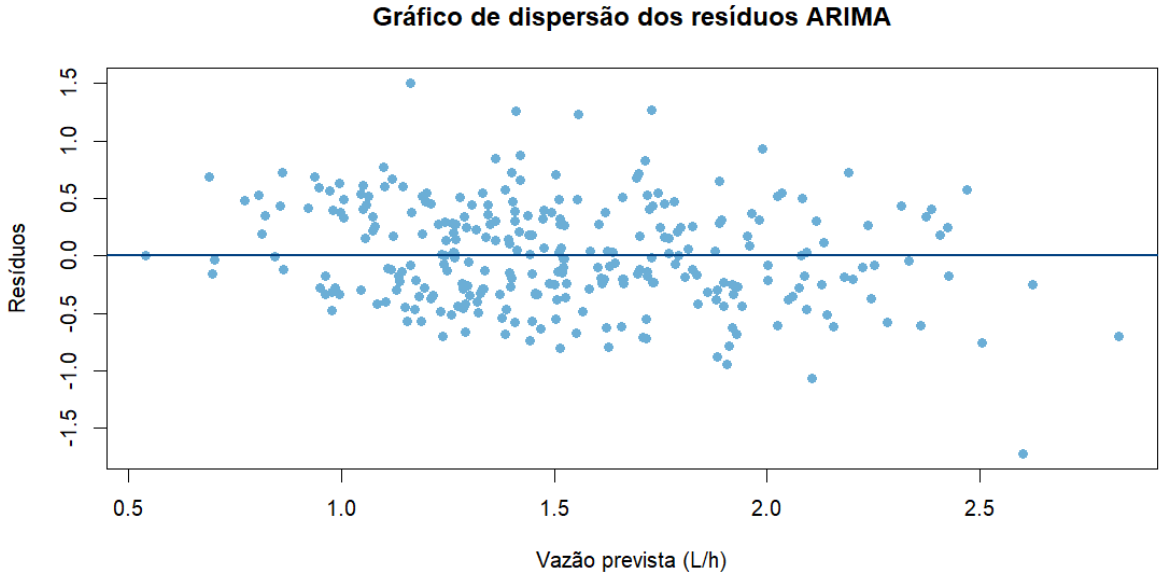


Figura 4.5 - Gráfico de dispersão dos resíduos ARIMA.

Nesse contexto, a análise gráfica dos resíduos revelou uma distribuição próxima à normalidade, sugerindo que os erros estão distribuídos de forma aleatória em torno da média zero. Além disso, observou-se que a variância dos resíduos permaneceu constante em relação aos valores previstos, indicando que o modelo também apresenta homocedasticidade.

Outrossim, foi realizado também o estudo de diferentes parâmetros no modelo para realizar um comparativo das métricas utilizadas em modelos mais simples, com

menos termos autorregressivos. A Tabela 4.7 demonstra a diferença entre os variados modelos.

Tabela 4.7 - Comparativo modelo ARIMA.

Modelo	AIC	MAE	RMSE	NSE	Valor-p do teste de Ljung-Box
ARIMA(1,1,0)	475,83	0,4277	0,5433	-0,0033	<2,2e-16
ARIMA(2,1,0)	424,38	0,4015	0,4953	0,1661	<2,2e-16
ARIMA(3,1,0)	419,89	0,3965	0,4898	0,1846	<2,2e-16
ARIMA(4,1,0)	420,58	0,3971	0,4887	0,1883	<2,2e-16
ARIMA(5,1,0)	**370,51**	**0,3644**	**0,4462**	**0,3234**	**2,265e-14**

Por conseguinte, observa-se a melhora de todas as métricas conforme o modelo se mostrou mais complexo, ou seja, com um maior parâmetro de autoregressividade, até chegar no modelo ARIMA(5,1,0).

Todavia, ainda com todas as limitações levantadas e identificadas, optou-se por continuar a previsão do consumo de água para os últimos 20% dos dados como parte do estudo, utilizando o pacote *forecast* do *RStudio*. A Tabela 4.8 apresenta as métricas de avaliação para a previsão gerada pelo modelo ARIMA(5,0,1).

Tabela 4.8 - Métricas de avaliação da previsão ARIMA.

MAE	RMSE	NSE
0,3792	0,5020	0,0176

Os valores dos erros calculados sugerem que, apesar das previsões geradas pelo modelo ARIMA, a sua precisão é bastante limitada. O valor baixo do NSE, em particular, destaca que o modelo não conseguiu oferecer uma melhoria significativa em relação ao simples uso da média dos valores observados, o que indica uma performance insatisfatória.

Essa baixa eficiência pode ser atribuída à ausência de consideração de sazonalidade no modelo. A série temporal de consumo de água apresenta padrões sazonais claros, e a falta de um componente sazonal na modelagem pode ter contribuído para a inadequação das previsões. A variabilidade nos padrões de consumo ao longo do tempo, muitas vezes influenciada por fatores, como variações

de temperatura e umidade, bem como por outros fatores externos, como mudanças na população ou eventos específicos, não foi adequadamente capturada.

Portanto, a limitação do modelo ARIMA, nesse contexto, ressalta a necessidade de incorporar fatores sazonais. Considerar modelos que lidem explicitamente com sazonalidade e variáveis externas pode proporcionar uma representação mais robusta e precisa dos padrões de consumo, possibilitando previsões mais confiáveis.

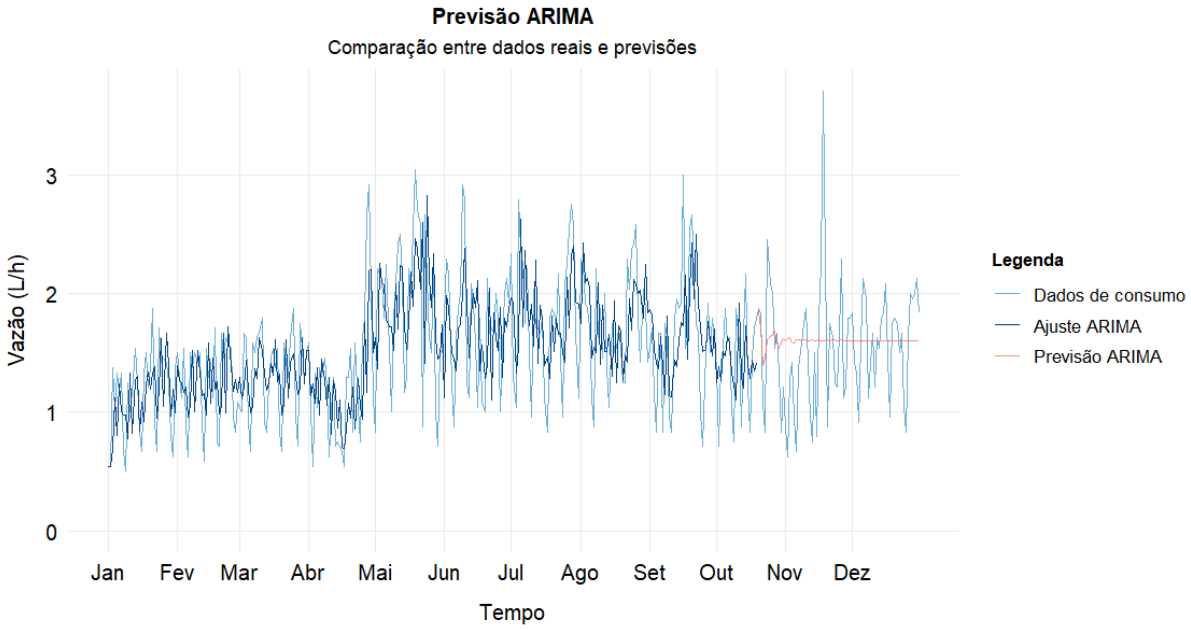


Figura 4.6 - Previsão ARIMA.

Além disso, ao observar o gráfico de previsão, demonstrado na Figura 4.6, observa-se que a previsão do modelo ARIMA resultou em previsões pouco assertivas a curto prazo seguidas de uma linha reta, que se aproxima da média de consumo para o período teste. Esse comportamento sugere que o modelo não conseguiu capturar as variações significativas nos dados de consumo de água, apresentando previsões quase que constantes ao longo do período analisado.

A presença de uma linha reta nas previsões indica que o modelo pode estar restringido em sua capacidade de adaptar-se às flutuações reais dos dados. Em um cenário ideal, o gráfico de previsões deveria mostrar variações que refletem os padrões históricos do consumo de água, incluindo picos e vales que são típicos em séries temporais complexas.

Ademais, foi possível identificar uma limitação significativa no pacote *forecast*, que se restringe ao uso exclusivo dos valores modelados para realizar previsões. No

caso específico do modelo ARIMA, constatou-se que, no curto prazo, o desempenho na previsão de valores futuros apresentou uma precisão razoável. Entretanto, à medida que o horizonte de previsão se estende, a eficácia do modelo diminui. Para melhorar a acurácia em previsões de longo prazo, seria necessário realimentar o modelo com novos dados observados ao longo do tempo, permitindo que ele capture variações recentes na série histórica e, assim, ajuste suas estimativas de forma mais robusta.

4.2.3 Modelo SARIMA

O modelo SARIMA selecionado, conforme identificado pelo *auto.arima*, foi SARIMA(1,0,1)(0,1,2)[7]. Os coeficientes estimados são apresentados na Tabela 4.9.

Tabela 4.9 - Coeficientes modelo SARIMA.

Coeficiente	Valor	Erro padrão	Valor-p
ar1	0,8620	0,0548	<2,2e-16
ma1	-0,3659	0,0960	0,0001389
sma1	-1,0364	0,0638	<2,2e-16
sma2	0,1191	0,0593	0,0446940

Os resultados indicam que todos os coeficientes são estatisticamente significativos, com valores de p menores que 0,05, o que demonstra a importância deles no modelo.

Desse modo, foram feitos testes do ajuste do modelo em relação aos valores observados, utilizando as métricas de avaliação AIC, MAE, RMSE e NSE, assim como detalhado na Tabela 4.6.

Tabela 4.10 - Métricas de avaliação do ajuste SARIMA.

AIC	MAE	RMSE	NSE
141,82	0,2158	0,2941	0,7061

O valor do Critério de Informação de Akaike sugere um equilíbrio entre a qualidade do ajuste e a sua simplicidade, mostrando uma menor perda de conhecimento em relação ao modelo ARIMA, logo, pelo AIC o modelo ajustado captura melhor os padrões de consumo observados. O MAE, por sua vez, se mostrou baixo, indicando também um ajuste preciso, com erros médios relativamente

pequenos. Já o RMSE, que penaliza erros maiores de forma mais severa em comparação com o MAE, foi próximo de 0,3, sugerindo que, apesar de existirem alguns erros de maior magnitude, o modelo ainda consegue modelar bem as flutuações observadas. Por fim o Nash-Sutcliffe de 0,7061 indicou um ajuste razoavelmente bom, capturando mais de 70% da variabilidade dos dados observados, que embora não seja perfeito, é eficaz para realizar previsões.

Os resultados indicam uma melhora no ajuste do modelo SARIMA em relação ao ARIMA, com um bom desempenho em termos de precisão e eficiência. Logo, cria-se uma base sólida para considerar o modelo como uma opção viável para entendimento e previsão de consumo de água.

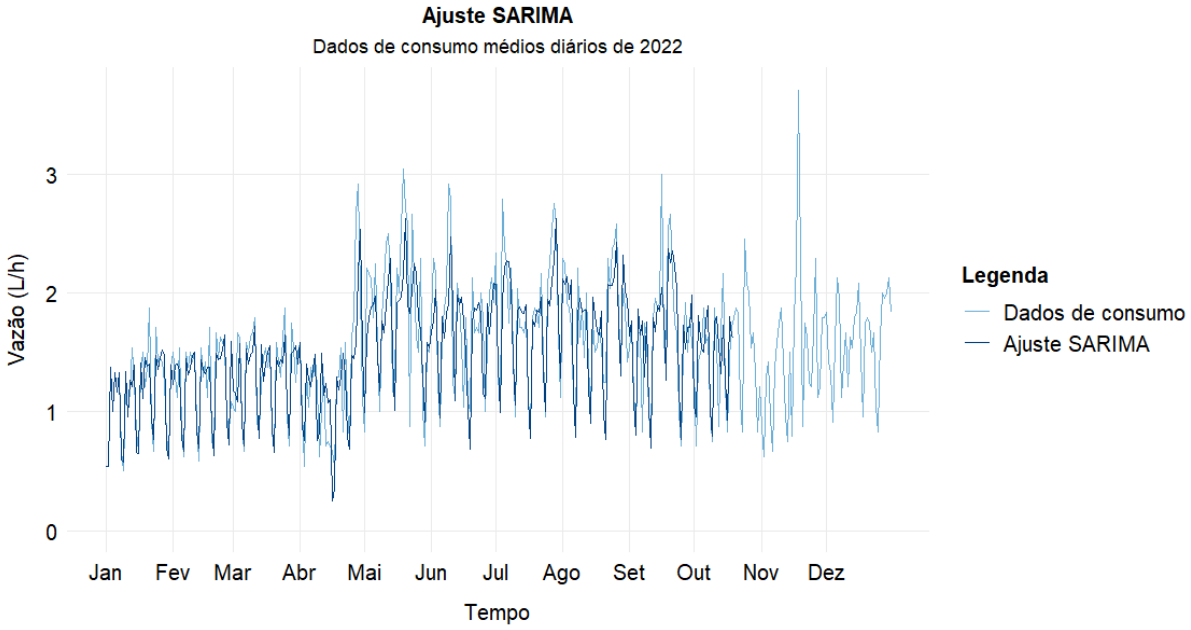


Figura 4.7 - Ajuste SARIMA.

Observando o gráfico na Figura 4.7, nota-se que o modelo ajustado conseguiu capturar bem a tendência da série temporal. No entanto, ele ainda apresenta limitações na previsão de grandes variações no consumo de água.

Em relação aos resíduos do modelo, o teste de *Ljung-Box* apresentou um valor-p de 0,8139. Esse resultado contribui para a rejeição da hipótese nula, indicando que os resíduos não são autocorrelacionados e se comportam como um ruído branco. Isso sugere que o modelo foi bem ajustado, pois os resíduos representam apenas a variabilidade aleatória dos dados, sem padrão de dependência temporal remanescente. Assim, também foi realizada a análise gráfica, que concluiu tanto a

normalidade quanto a homoscedasticidade dos resíduos. Os gráficos plotados podem ser visualizados nas Figura 4.8 e Figura 4.9.

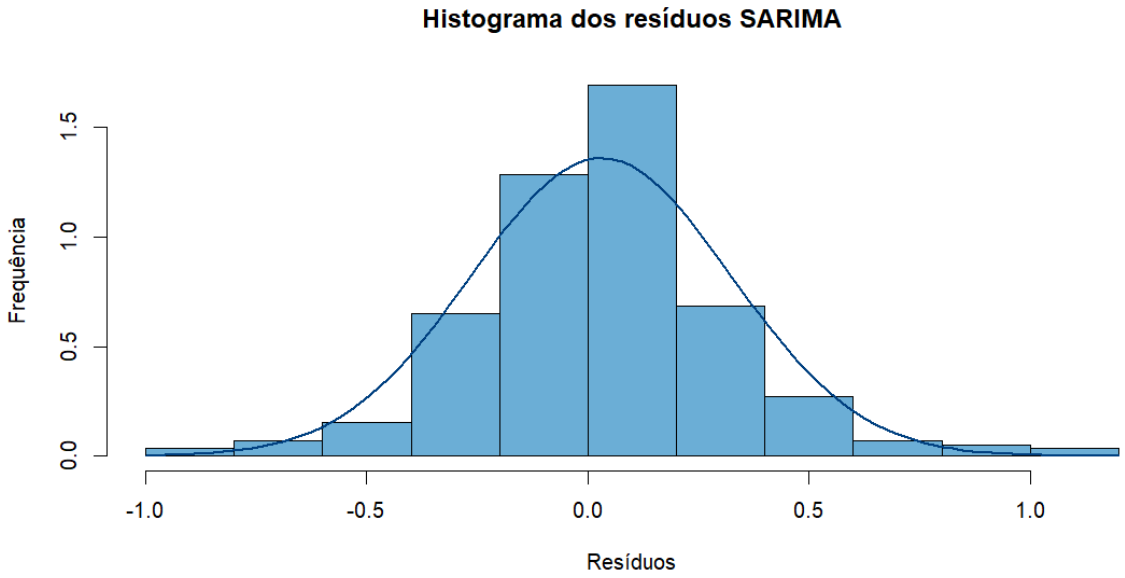


Figura 4.8 - Histograma dos resíduos SARIMA.

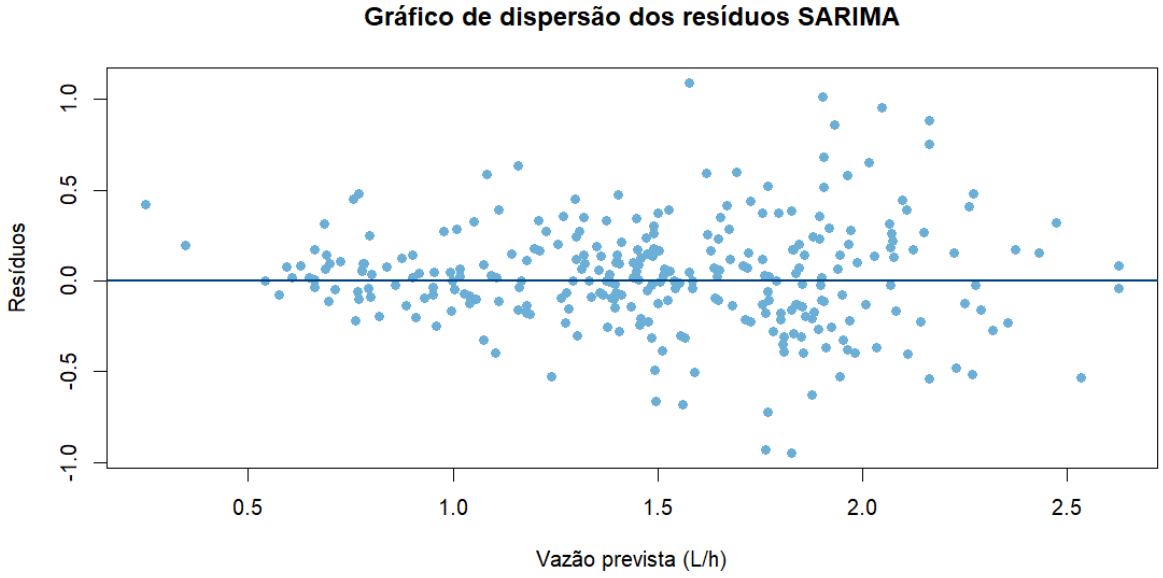


Figura 4.9 - Gráfico de dispersão dos resíduos SARIMA.

Da mesma forma como realizado para o modelo ARIMA, também foram estudados diferentes modelos como critério comparativo, tal como mostra a Tabela 4.11. Isto posto, observa-se que modelos mais simples mostram métricas piores,

enquanto modelos mais complexos mesmo apresentam erros menores, são penalizados no critério de *Akaike*, pela maior quantidade de parâmetros.

Tabela 4.11 - Comparativo modelos SARIMA.

Modelo	AIC	MAE	RMSE	NSE	Valor-p do teste de Ljung-Box
SARIMA(1,0,1)(0,1,1)[7]	143,63	0,2177	0,2955	0,7031	0,489
SARIMA(1,0,1)(0,1,2)[7]	**141,82**	**0,2158**	**0,2941**	**0,7061**	**0,814**
SARIMA(2,0,2)(0,1,2)[7]	142,87	0,2147	0,2919	0,2719	0,841
SARIMA(2,0,3)(0,1,3)[7]	146,38	0,2143	0,2917	0,7107	0,863

Consequentemente, prosseguiu-se com a previsão do modelo utilizando a série temporal de teste, com o objetivo de verificar se a qualidade da previsão seria comparável à do ajuste inicial.

Tabela 4.12 - Métricas de avaliação da previsão ARIMA.

MAE	RMSE	NSE
0,3372	0,4656	0,2098

O Erro Médio Absoluto (MAE) de 0,3372 sugere que, em média, o modelo apresenta um desvio de 0,3372 unidades em relação aos valores observados, o que aponta para uma precisão razoável na previsão. No entanto, isso também evidencia algumas limitações na captura e modelagem das variações nos dados de consumo de água. O valor do Erro Quadrático Médio (RMSE) de 0,4656 corrobora essa observação, indicando que o modelo enfrenta dificuldades em prever grandes flutuações no consumo de água, o que resulta em discrepâncias mais acentuadas para determinados pontos de dados.

Por fim, o Coeficiente de Eficiência de Nash-Sutcliffe (NSE) de 0,2098 demonstra que o modelo SARIMA não está capturando adequadamente a variabilidade do consumo de água no período teste. Esse valor sugere que, utilizar o modelo ajustado como previsão poderia ser uma abordagem mais eficaz do que a própria média. Entretanto, ainda assim, o modelo não conseguiu reproduzir a

complexidade do comportamento do consumo de água na Península Norte de forma satisfatória, indicando a necessidade de melhorias na metodologia de previsão ou até mesmo nos dados de entrada.

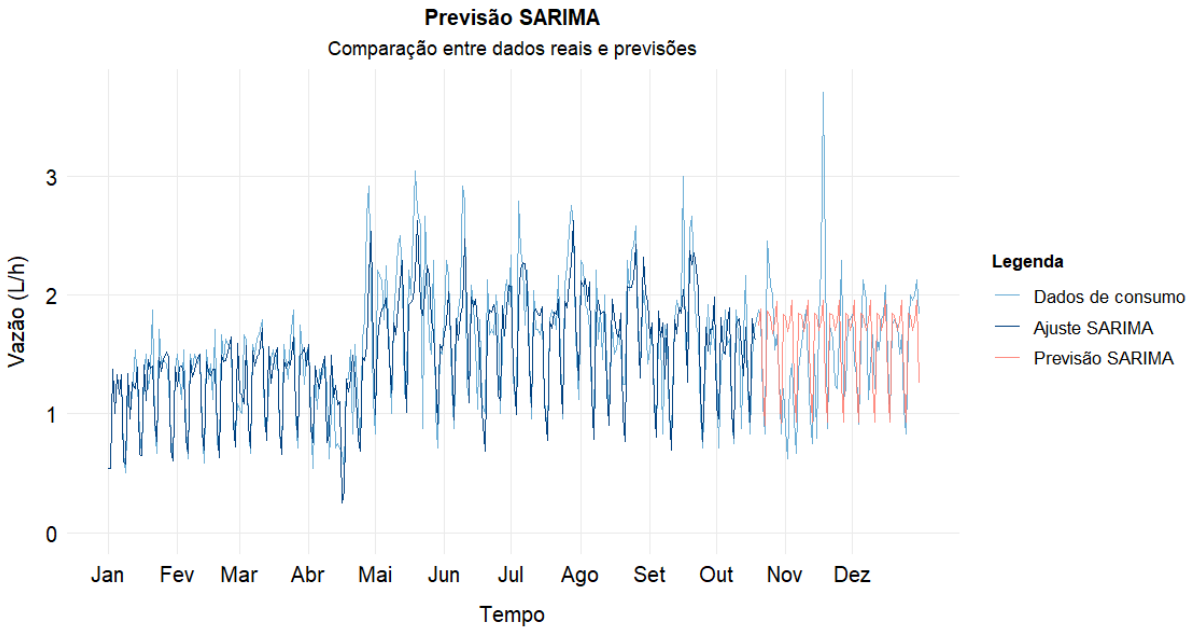


Figura 4.10 - Previsão SARIMA.

O gráfico de previsão, apresentado na Figura 4.10, indica uma melhoria em relação ao modelo ARIMA, evidenciando que a previsão não se estabiliza rapidamente e acompanha um período mais longo da série temporal.

A previsão obtida reflete uma sazonalidade semanal, capturando adequadamente os padrões de consumo de água que se repetem a cada 7 dias. Contudo, o modelo apresenta limitações significativas ao prever picos e quedas de consumo durante o período de teste. Embora o modelo demonstre uma leve tendência de aumento, essa tendência é mínima e mantém o mesmo padrão semanal ciclicamente. Isso sugere que os dados utilizados não são suficientemente representativos para captar a complexidade total do comportamento de consumo, que exibe uma variabilidade considerável.

Ainda assim, o modelo SARIMA demonstra uma capacidade de previsão a curto prazo. Apesar de ele se limitar na previsão de flutuações mais agudas, é possível considerá-lo eficaz para identificar tendências gerais e padrões sazonais. De maneira semelhante ao observado no modelo ARIMA, seria ideal que o modelo fosse continuamente retroalimentado com novos dados à medida que estes se tornassem disponíveis. Esse processo de atualização permitiria ajustar o modelo às mudanças

mais recentes da série temporal, resultando em previsões mais precisas e eficientes ao longo do tempo.

4.2.4 Modelo de Regressão Linear Multivariada

Para o modelo de regressão linear multivariada, foram definidas as variáveis explicativas com base na vazão de consumo média iniciada naquele dia, a vazão do dia anterior e as vazões dos mesmos dias das semanas nas 7 semanas anteriores. As variáveis independentes podem ser observadas na Tabela 4.13.

Tabela 4.13 - Variáveis independentes do modelo de regressão linear múltipla.

Variável explicativa	Código
Vazão média de consumo iniciada naquele dia	Qm
Vazão de consumo no dia anterior	Q_dia_anterior
Vazão de consumo no mesmo dia da semana anterior	Q_semana_1
Vazão de consumo no mesmo dia de 2 semanas anteriores	Q_semana_2
Vazão de consumo no mesmo dia de 3 semanas anteriores	Q_semana_3
Vazão de consumo no mesmo dia de 4 semanas anteriores	Q_semana_4
Vazão de consumo no mesmo dia de 5 semanas anteriores	Q_semana_5
Vazão de consumo no mesmo dia de 6 semanas anteriores	Q_semana_6
Vazão de consumo no mesmo dia de 7 semanas anteriores	Q_semana_7

Isto posto, para a seleção das variáveis significativos, foi construída uma matriz de correlação de Pearson com todas as variáveis explicativas desejadas, de tal forma, que com o resultado apresentado nas Figura 4.11 e Figura 4.12 e na Tabela 4.14, foi

possível selecionar as variáveis relevantes para a modelagem de consumo de água no presente estudo de caso.

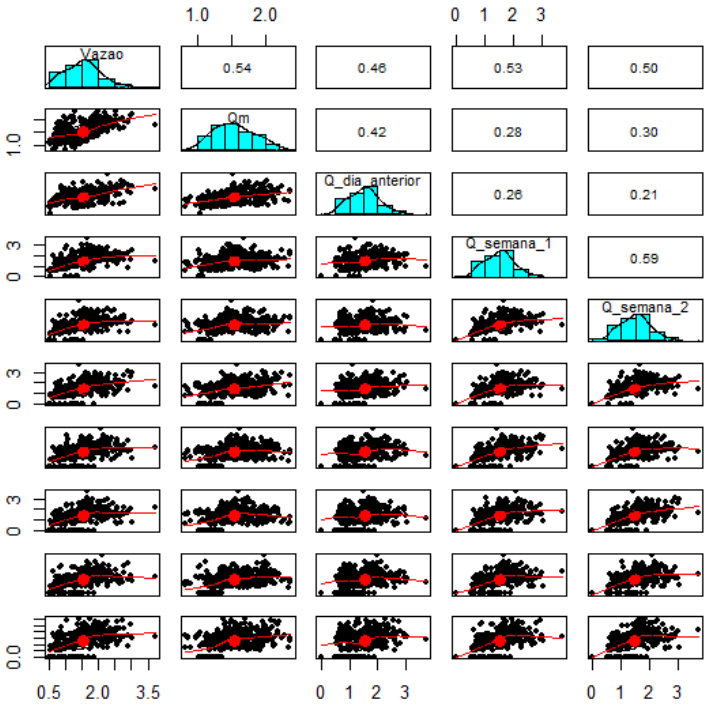


Figura 4.11 - Matriz de correlação de Pearson para as variáveis independentes - Parte 1.

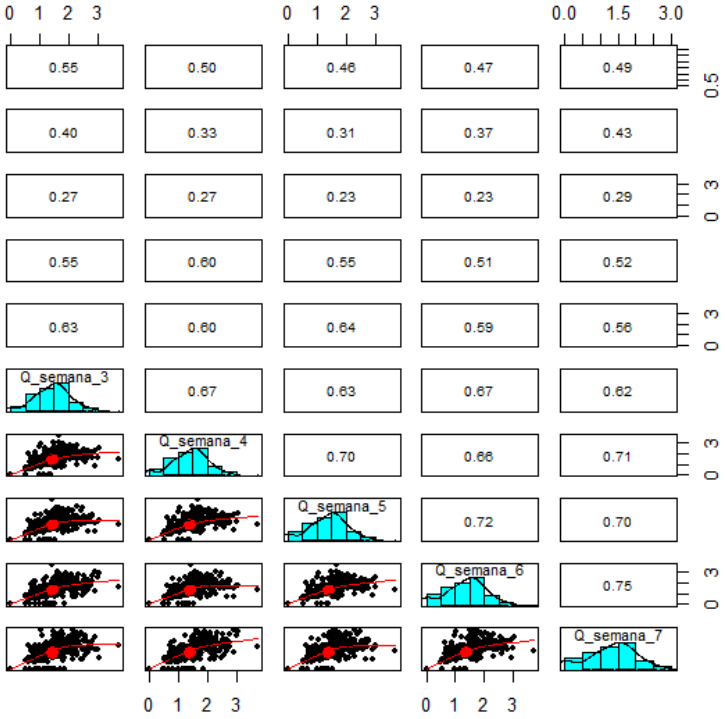


Figura 4.12 - Matriz de correlação de Pearson para as variáveis independentes - Parte 2.

Tabela 4.14 - Coeficientes estimados para as variáveis explicativas.

Coeficiente	Valor estimado	Erro padrão	Valor-p
Qm	0,4635	0,0753	2,02e-09
Q_dia_anterior	0,2103	0,0405	3,52e-07
Q_semana_1	0,1980	0,0465	2,64e-05
Q_semana_2	0,0935	0,0475	0,0497
Q_semana_3	0,1388	0,0484	0,0044
Q_semana_4	0,0338	0,0495	0,4951
Q_semana_5	-0,0012	0,0479	0,9794
Q_semana_6	0,0090	0,0479	0,8508
Q_semana_7	-0,0068	0,0475	0,8859

A partir dos resultados observados, notou-se que os valores-p para as variáveis a partir da terceira semana anterior (Q_semana_3) foram superiores a 0,05, indicando falta de significância desses coeficientes para o modelo ajustado, logo, as demais variáveis com valores-p bem acima do limite foram retiradas do modelo.

Desse modo, o modelo final utilizou as seguintes variáveis significativas, cujos coeficientes estão apresentados na Tabela 4.15.

Tabela 4.15 - Coeficientes estimados para as variáveis significativas.

Coeficiente	Valor estimado	Erro padrão	Valor-p
Qm	0,4701	0,0732	4,35e-10
Q_dia_anterior	0,2128	0,0402	2,09e-07
Q_semana_1	0,2106	0,0442	2,76e-06
Q_semana_2	0,1037	0,0446	0,0204
Q_semana_3	0,1588	0,0426	0,0002

Foi possível verificar que todos os valores-p estabelecidos nesse modelo apresentaram valor inferior a 0,05, indicando que esses parâmetros são estatisticamente significativos ao modelo ajustado. Dessa forma, a matriz de correlação de Pearson se apresentou como segue na Figura 4.13.

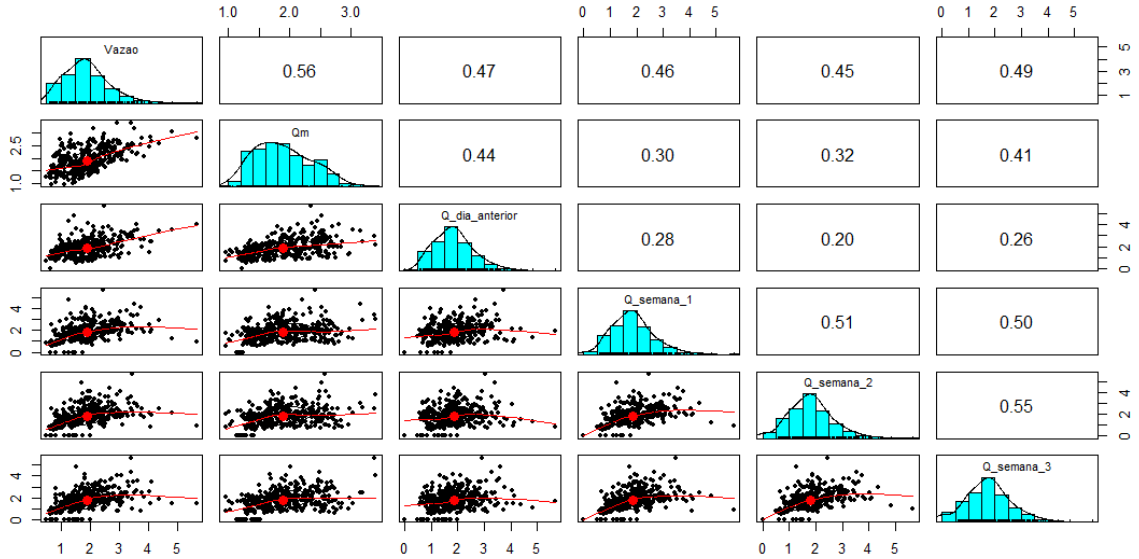


Figura 4.13 - Matriz de correlação de Pearson para as variáveis significativas.

Nesse ínterim, os valores de multicolinearidade, situação em que uma ou mais variáveis explicativas apresentam autocorrelação, indicam de média para baixa correlação entre as variáveis independentes, sugerindo que o modelo possui boa qualidade. Portanto, o modelo de regressão linear multivariada resultou na seguinte formulação para previsão de consumo de água:

$$Q = 0{,}4701 * Qm + 0{,}2128 * Q_{dia_{anterior}} + 0{,}2106 * Q_{semana_1} + 0{,}1037 * Q_{semana_2} + 0{,}1588 * Q_{semana_3} \quad (4.1)$$

Tabela 4.16 - Métricas de avaliação do modelo de regressão linear multivariada.

AIC	R^2	RSE
319,983	0,5213	0,371

Nessa perspectiva, observa-se que o modelo de regressão linear multivariada se mostrou intermediário ao modelo SARIMA e ARIMA, uma vez que apresentou o critério de conhecimento de Akaike em 319,983 devido a sua maior complexidade quando comparado a ambos os métodos e o seu pior ajuste, quando comparado com o método SARIMA.

Além disso, o coeficiente de determinação ajustado foi de 0,5213, indicando que 52,13% da variabilidade de vazão de consumo de água é capturada pelo modelo, isto é, 52,13% podem ser explicadas pelas variáveis independentes incluídas no modelo. Esse valor indica que o modelo tem uma capacidade média de explicar os dados observados, uma vez que um pouco menos de 50% do modelo permanece não explicada, sugerindo que pode haver outras variáveis ou fatores que não foram considerados que influem significativamente a variável dependente.

O erro residual padrão, que reflete a dispersão dos resíduos em relação ao modelo de regressão ajustada, foi de 0,371, indicando que os dados se apresentaram razoavelmente próximos à modelagem da regressão.

Esses resultados demonstram que a regressão linear multivariada foi eficaz em capturar a relação entre as variáveis explicativas e a vazão de água, tal que ainda há espaço para melhorias, mas evidenciando que a vazão média, a vazão do dia anterior e a vazão de semanas anteriores exercem influência significativa sobre as previsões de consumo.

5 CONCLUSÃO E RECOMENDAÇÕES

A análise dos modelos de previsão de consumo de água realizados neste estudo, que incluiu o ARIMA, SARIMA e a Regressão Linear Multivariada, forneceu informações valiosas sobre a eficácia dessas abordagens estatísticas para captar e prever a dinâmica do consumo de água na Península Norte e em regiões similares. A comparação entre os modelos evidenciou as forças e limitações de cada um, destacando áreas cruciais que requerem investigação adicional para aprimorar a gestão de recursos hídricos e a precisão das previsões, como a sazonalidade a ser utilizada, a quantidade de dados disponíveis para o estudo e as variáveis explicativas adicionadas ao modelo.

Todavia, também é importante destacar que a qualidade dos dados desempenha um papel central na eficácia dos modelos de previsão. No presente estudo, verificou-se que uma parte significativa dos dados disponíveis apresentou falhas ou inconsistências que poderiam ter sido tratadas com mais rigor. Dados de qualidade superior, abrangendo períodos mais longos, como a partir de 2021, permitiriam a modelagem de padrões anuais e uma análise mais profunda da evolução do consumo ao longo dos anos, potencialmente revelando tendências de longo prazo. Modelos que incluíssem um horizonte temporal maior teriam a capacidade de captar melhor os efeitos sazonais do consumo de água na região.

Igualmente importante é o processo de tratamento dos dados. Dados brutos frequentemente apresentam problemas como valores ausentes, *outliers* e *spikes* decorrentes de falhas em sistemas de medição, como já destacado em outros trechos. Sem um tratamento rigoroso, como a aplicação de métodos estatísticos adequados para lidar com esses problemas, os modelos podem ser enviesados, comprometendo a confiabilidade das previsões. A aplicação de técnicas de limpeza, como a utilização do método IQR para lidar com *outliers*, e a interpolação para dados ausentes, são essenciais para garantir que os dados representem com precisão o comportamento real do consumo de água. A qualidade e o tratamento eficaz dos dados são, portanto, fundamentais para construir modelos estatísticos robustos, reduzindo incertezas e melhorando a confiabilidade dos resultados obtidos.

Isto posto, o modelo ARIMA demonstrou ser insuficiente para captar os padrões complexos e sazonais do consumo de água, como indicado pelos baixos valores do Coeficiente de Eficiência de Nash-Sutcliffe (NSE). Esses resultados sugerem que as previsões baseadas na média poderiam ser tão eficazes quanto o próprio modelo.

Além disso, a análise dos resíduos, que apresentam comportamento normal e homocedástico, indicou que eles não se comportaram como ruído branco, revelando que o modelo falhou em compreender as características intrínsecas da série temporal de consumo de água. Estudos, como o de Tsutyia (2004) e de outros pesquisadores, destacam que o consumo de água apresenta variações influenciadas por fatores como clima, estações do ano, hábitos populacionais e temperatura, fato esse que, o modelo ARIMA não conseguiu capturar possivelmente devido à ausência de uma componente sazonal ou até mesmo da autocorrelação de longo período, hipóteses que dependem de uma maior gama de dados para conclusão exata.

O modelo SARIMA, ao incorporar uma componente sazonal, mostrou melhorias significativas em relação ao ARIMA. O AIC menor indicou um ajuste mais adequado aos dados, e as métricas de erro sugeriram que o SARIMA conseguiu captar melhor a variabilidade do consumo. Além disso, os resíduos apresentaram características de ruído branco, comportamento normal e homoscedasticidade confirmando um ajuste mais robusto e apropriado. No entanto, o SARIMA ainda enfrentou limitações na previsão de grandes variações de consumo, logo, embora o ajuste do modelo tenha sido satisfatório, a previsão de picos e vales de consumo permaneceu como um desafio. Essa limitação pode estar relacionada à falta de dados adicionais que poderiam capturar padrões de consumo mais complexos. No presente estudo foi utilizada uma sazonalidade semanal, todavia, observou-se que os padrões de consumo de água podem ser influenciados por eventos anuais específicos, como feriados e datas comemorativas, variáveis climáticas e entre outras, sugerindo que um modelo que abranja sazonalidade anual poderia proporcionar resultados de previsão mais precisos e abrangentes.

A regressão linear multivariada, que incorporou variáveis explicativas relacionadas à vazão média de consumo e às vazões de dias anteriores, mostrou-se uma abordagem promissora para a previsão de consumo de água. A seleção criteriosa de variáveis significativas resultou em coeficientes estatisticamente robustos, indicando que a regressão linear multivariada pode oferecer previsões úteis, especialmente quando combinada com considerações de fatores sazonais e outras variáveis explicativas, como temperatura e umidade relativa do ar. Embora o modelo de regressão linear multivariada não tenha apresentado os melhores resultados no presente estudo, seu potencial para prever o consumo de água é evidente, desde que

sejam consideradas outras variáveis que influenciam o consumo e geram as flutuações observadas nos dados.

Tabela 5.1 - Comparativo entre modelos.

Modelo	**Ajuste**			
	AIC	**MAE**	**RMSE**	**NSE**
ARIMA	370,51	0,3643	0,4462	0,3234
SARIMA	141,82	0,2158	0,2941	0,7061
Regressão Linear Multivariada	**AIC**	**RSE**		**R^2**
	319,983	0,371		0,5213

Nessa perspectiva, conclui-se que o modelo ARIMA, por si só, mostrou-se insuficiente para prever com precisão os valores de consumo de água, especialmente em séries temporais complexas. O modelo SARIMA, por outro lado, demonstrou maior potencial de previsão, desde que haja uma maior disponibilidade de dados, tanto em quantidade quanto em qualidade, o que possibilitaria captar melhor as flutuações sazonais e os padrões de consumo. Por fim, o modelo de Regressão Linear Multivariada também se revelou promissor, mas seu desempenho poderia ser significativamente aprimorado com a inclusão de variáveis independentes adicionais, que influenciam diretamente o comportamento da demanda por água.

Outrossim, conforme apontado na dissertação de Costa (2022), todos os modelos apresentaram resultados satisfatórios na análise direta das unidades consumidoras, demonstrando sua funcionalidade com o uso de dados de telemetria. No entanto, foi observado que, à medida que aumentava a variabilidade do consumo e a quantidade de leituras com ausência de consumo, os resultados dos modelos se tornavam menos precisos. Um ponto importante a ser destacado é que o estudo de Costa (2022) utilizou dados provenientes de uma amostra de 12 dos 423 hidrômetros da área de estudo (Quadra 12 do Setor Habitacional Jardins Mangueiral), os quais transmitiam os valores de consumo a cada 4 horas.

Diante disso, observa-se que, para unidades consumidoras autônomas, os modelos propostos apresentaram um desempenho favorável, conforme concluído por Costa (2022). Contudo, ao comparar os estudos, observa-se que para o presente trabalho foram obtidos melhores ajustes e previsão para os modelos, mesmo que para

uma análise mais abrangente do consumo de uma região inteira, seria necessário implementar algumas adaptações e melhorias. Porém ainda sim é perceptível os pontos comuns entre os estudos, tal que ambos concluíram a falta da adaptabilidade dos modelos com elevadas variações de consumo. Somente com essas mudanças os modelos poderiam ser aplicados de maneira mais eficaz e precisa, garantindo maior confiabilidade nas previsões e no gerenciamento dos recursos hídricos em larga escala.

Nessa perspectiva, recomenda-se para futuros estudos:

- Incorporação de dados sazonais mais detalhados para melhorar a precisão das previsões, recomenda-se a utilização de modelos que considerem múltiplas sazonalidades, incluindo padrões diários, semanais, mensais e anuais. Esses modelos podem capturar de forma mais precisa as variações complexas do consumo de água, especialmente em regiões onde o consumo varia significativamente ao longo do ano.
- Coleta e análise de dados adicionais: a obtenção de uma gama maior de dados de telemetria é crucial. Aumentar o período de coleta de dados permitirá a detecção de padrões de longo prazo e sazonais anuais, que são fundamentais para entender as variações de consumo associadas a eventos específicos e mudanças climáticas.
- Incorporação de variáveis externas: a inclusão de variáveis externas, como dados climáticos (temperatura, precipitação), eventos sociais e econômicos, e mudanças populacionais, pode melhorar substancialmente a capacidade de previsão dos modelos. Essas variáveis são frequentemente correlacionadas com o consumo de água e podem ajudar a capturar flutuações que os modelos baseados apenas em dados históricos de consumo não conseguem prever.

Por fim, investir em métodos de previsão aprimorados e na coleta de dados mais detalhados será essencial para melhorar a gestão de recursos hídricos, permitindo a formulação de políticas mais eficazes e sustentáveis para o uso da água. O avanço contínuo na modelagem e a integração de novos métodos e dados são fundamentais para enfrentar os desafios da gestão de recursos hídricos em um mundo cada vez mais urbanizado e afetado pelas mudanças climáticas.

6 BIBLIOGRAFIA

ABNT. **NBR 8009 - Hidrômetro taquimétrico para água fria até 15,0 m³/h de vazão nominal**. Rio de Janeiro: Associação Brasileira de Normas Técnicas, 1997.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716-723, 1974.

ALEGRE, H. et al. **Controlo de perdas de água em sistemas públicos de adução e distribuição**. [S.l.]: Instituto Regulador de Águas e Resíduos, 2005.

ANELE, A. O. et al. Overview, Comparative Assessment and Recommendations of Forecasting Models for Short-Term Water Demand Prediction. **Water**, v. 9, n. 5, p. 409, 2017.

BANDEIRA, K. Brasília cresce de forma desordenada e vira 3ª maior cidade do país. **O Globo**, 2023. Disponivel em: <https://oglobo.globo.com/brasil/noticia/2023/07/16/planejada-no-passado-brasilia-cresce-de-forma-desordenada-e-vira-3a-maior-cidade-do-pais.ghtml>. Acesso em: 2024 set. 16.

BRANDÃO, A.; PAVIANI, A. **A crise e o consumo de água em Brasília**. Brasília: Editora UnB, 2018.

BRASIL. Lei nº 11.445, de 5 de janeiro de 2007. Estabelece as diretrizes nacionais para o saneamento básico. **Diário Oficial da União**, Brasília, DF, 8 jan. 2007.

BRASIL. Lei nº 14.026, de 15 de julho de 2020. Atualiza o marco legal do saneamento básico. **Diário Oficial da União**, Brasília, DF, 15 jul. 2020.

CAESB. Lago Norte é a primeira cidade a receber o sistema de telemetria. **Caesb DF**, 2019. Disponivel em: <https://www.caesb.df.gov.br/8-noticias/993-27-11-19-lago-norte-e-a-primeira-cidade-a-receber-o-sistema-de-telemetria.html>. Acesso em: 21 nov. 2023.

CAESB. **Plano diretor de água e esgotos do Distrito Federal**. Brasília: Compainha de Saneamento Ambiental do Distrito Federal, 2019.

CODEPLAN. **Pesquisa Distrital por Amostra de Domicílios**. Brasília: Compainha de Planejamento do Distrito Federal, 2021.

COSTA, D. F. **Avaliação dos Benefícios de Medição por Telemetria em Rede de Água:** Estudo, Modelagem do Consumo e Controle de Perdas. Dissertação (Mestrado em Tecnologia Ambiental e Recursos Hídricos). Brasília: Universidade de Brasília. 2022.

DE CASTRO, K. B. **Consumo de água do Distrito Federal por Região Administrativa**. Brasília: Compainha de Planejamento do Distrito Federal, 2018.

DIAO, K.; BARJENBRUCH, M.; BRACKLOW, U. Study on The Impacts of Peaking Factor on a Water Distribution System in Germany. **Water Science and Technology: Water Supply**, v. 10, p. 165-172, 2010.

DONKOR, E. A. et al. Urban water demand forecasting: Review of methods and models. **Journal of Water Resources Planning and Management**, v. 138, n. 2, p. 146-159, 2012.

FILHO, N. A. C. **Avaliação de medidores com múltiplas tomadas de pressão na macromedição em sistemas de água**. Ilha Solteira: Universidade Estadual Paulista, 2007.

FILHO, N. A. C.; DALL'AGLIO, M. Avaliação de medidores de vazão com múltiplas tomadas de pressão na macromedição em sistemas de abastecimento de água. **Revista DAE**, n. 178, p. 40-47, 2006.

GDF. **Plano Distrital de Saneamento Básico**. Brasília: Governo do Distrito Federal, 2017.

GUIDI, J. M. M. Análise da variação do consumo de água utilizando dados obtidos por sistemas supervisores remotos. Estudo de caso: Município de Franca - SP. **Congresso ABES**, São Paulo, p. 15, 2017.

HYDRUS. **DIEHL Metering Ultrasonic Meter**. Lisle: Diehl Metering LLC, 2024.

IBGE. **Cidades**, 2022. Disponivel em: <https://cidades.ibge.gov.br/brasil/df/panorama>. Acesso em: 03 set. 2024.

INMETRO. **Portaria nº 155**. Brasília: INMETRO, 2022.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Cidades**, 2022. ISSN https://cidades.ibge.gov.br/brasil/df/panorama. Acesso em: 03 set. 2024.

LIMA, C. A. F.; NAVAS, J. R. P. Smart metering and systems to support a conscious use of water and eletricity. **Energy**, Campinas, v. 45, p. 528-540, 2012.

ONU. **Resolução 64/292**. [S.l.]: [s.n.], 2010.

ONU. **Relatório Mundial das Cidades**. ONU Habitat. [S.l.]. 2020.

PASTEGA, G. Outliers: Como definir, detectar e tratar — Parte 2. **Medium**, 2023. Disponivel em: <https://medium.com/@gabrielpbreis/outliers-como-definir-detectar-e-tratar-parte-2-5240149f8f98>. Acesso em: 12 mar. 2024.

SECRETÁRIA NACIONAL DE SANEAMENTO AMBIENTAL. Diagnóstico Temático - Serviços de Água e Esgoto. **Ministério das Cidades**, Brasília, 2022.

SILVA, N. R. D. **Estudo de metodologias para avaliação de submedição de hidrômetros domiciliares em sistemas de água**. Brasília: [s.n.], 2008.

TRICARICO, C. et al. **Peak Residential Water Demand. Proceedings of the Institution of Civil Engineers:** Water Management. [S.l.]: [s.n.], 2007.

TSUTYIA, M. T. **Abastecimento de Água**. São Paulo: Departamento de Engenharia Hidráulica e Sanitária da Escola Politécnica da Universidade de São Paulo, 2004.

USP. **PSI 2461 - Eletrônica de controle industrial - Medidores de vazão**. São Paulo: [s.n.], 2017.

7 ANEXOS

7.1 Rotina de Tratamento de Dados no RStudio

```
# Projeto Final 2 - Enzo D'Angelo Arruda Duarte

# CONFIGURAÇÃO DO AMBIENTE
# Limpeza do ambiente
rm(list = ls())

# Carregar pacotes necessários
pacotes <- c("dplyr", "data.table", "stringr", "zoo", "tidyr","installr",
"lubridate","beepr")
for (pacote in pacotes) {
  if (!require(pacote, character.only = TRUE)) {
    install.packages(pacote)
    library(pacote, character.only = TRUE)
  }
}

# Definir o diretório de trabalho
setwd("C:/Users/enzod/OneDrive - unb.br/PF2/Dados")

# Atualizar o R
updateR()

# BASE DE DADOS
# Importação da base de dados
df <- fread("Leituras 2022.csv", header = FALSE, encoding = "UTF-8")

# TRATAMENTO INICIAL
# Deletar colunas indesejadas
df <- df[, -c(4, 6, 8), with = FALSE]

# Renomear colunas
setnames(df, old = c("V1", "V2", "V3", "V5", "V7"), new = c("ID", "Data",
"Leitura", "Equipamento", "Imovel"))

# Mudança das unidades dos dados
df[, ID := as.numeric(ID)]
df[, Imovel := as.integer(Imovel)]
df[, Data := as.POSIXct(Data, format = "%Y-%m-%d %H:%M:%S")]
df[, Leitura := as.numeric(gsub(",", ".", Leitura))]
```

```
# LIMPEZA INICIAL
# Remover duplicatas
df <- unique(df)

# Remover NAs
df <- na.omit(df)

# Retirar valores de leitura negativos e NAs
df <- df[!is.na(Leitura) & Leitura >= 0]

# Filtrar por ID do equipamento
df <- df[str_starts(Equipamento, "70B3")]

# VERIFICAÇÃO INICIAL
head(df)
summary(df)
str(df)

# TRATAMENTO DOS DADOS
# Ordenar os dados por Imovel e Data
setorder(df, Imovel, Data)

# AMOSTRA
# Função para calcular o tamanho da amostra
calcular_tamanho_amostra   <-   function(confidence_level,   error_margin,
total_population) {
  # Determinar o Z-score com base no nível de confiança
  z_score <- qnorm(1 - (1 - confidence_level) / 2)

  # Calcular o tamanho da amostra
  n_0 <- (z_score^2 * 0.5 * 0.5) / (error_margin^2)

  # Ajuste do tamanho da amostra para população finita
  n <- ceiling(n_0 / (1 + (n_0) / total_population))

  return(n)
}

# Parâmetros
nivel_confianca <- 0.95  # 95% de confiança
```

```
margem_erro <- 0.05      # 5% de margem de erro

# Número total de imóveis no dataframe
total_imoveis <- length(unique(df$Imovel))

# Calcular o tamanho da amostra
tamanho_amostra  <-  calcular_tamanho_amostra(nivel_confianca,  margem_erro,
total_imoveis)

# Contar o número de registros para cada imóvel
contagem_imoveis <- df %>%
  group_by(Imovel) %>%
  summarize(Contagem = n()) %>%
  arrange(desc(Contagem))

# Selecionar os imóveis com a maior quantidade de dados até atingir o tamanho
da amostra
imoveis_amostra <- contagem_imoveis %>%
  slice_head(n = tamanho_amostra) %>%
  pull(Imovel)

# Filtrar o dataframe original para incluir apenas esses imóveis selecionados
df_amostra <- df %>%
  filter(Imovel %in% imoveis_amostra)

# Exibir o resultado
cat("O tamanho da amostra necessário é:", tamanho_amostra, "\n")

# Criar intervalos de 4 em 4 horas dentro de cada dia
df_amostra[,   Intervalo_4h   :=   cut(as.POSIXct(format(Data,   "%H:%M:%S"),
format="%H:%M:%S"),
                         breaks         =         seq(as.POSIXct("00:00:00",
format="%H:%M:%S"),
                                            as.POSIXct("24:00:00",
format="%H:%M:%S"), by="4 hours"),
                         labels  =  c("00:00",  "04:00",  "08:00",  "12:00",
"16:00", "20:00"),
                         include.lowest = TRUE, right = FALSE)]

df_amostra[, Intervalo_4h := paste(format(Data, "%Y-%m-%d"), Intervalo_4h)]
# Concatenar a data com o intervalo de horas
```

```
# Identificar spikes na Leitura
df_amostra <- df_amostra %>%
  group_by(Imovel) %>%
  mutate(
    Q1 = quantile(Leitura, 0.25, na.rm = TRUE),
    Q3 = quantile(Leitura, 0.75, na.rm = TRUE),
    IQR = Q3 - Q1,
    Limite_Inferior = pmax(Q1 - 1.5 * IQR, 0), # Garante que o limite inferior
não seja menor que zero
    Limite_Superior = Q3 + 1.5 * IQR,
    Leitura = ifelse(Leitura < Limite_Inferior | Leitura > Limite_Superior,
NA, Leitura)
  ) %>%
  ungroup()

# Função para preencher NAs com progressão aritmética
preencher_NAs_progressao_aritmetica <- function(x) {
  # Encontrar os índices dos NAs
  na_indices <- which(is.na(x))

  # Se não há NAs, retornar o vetor original
  if (length(na_indices) == 0) {
    return(x)
  }

  # Vetor para armazenar o resultado
  resultado <- x

  # Iterar por cada bloco de NAs
  for (i in seq_along(na_indices)) {
    indice_na <- na_indices[i]

    # Encontrar o início do bloco de NAs
    if (i == 1 || na_indices[i] != na_indices[i - 1] + 1) {
      bloco_inicio <- indice_na
    }

    # Encontrar o fim do bloco de NAs
    if (i == length(na_indices) || na_indices[i] != na_indices[i + 1] - 1) {
      bloco_fim <- indice_na
```

```
      # Valores antes e depois do bloco de NAs
      valor_antes <- if (bloco_inicio > 1) x[bloco_inicio - 1] else NA
      valor_depois <- if (bloco_fim < length(x)) x[bloco_fim + 1] else NA

      # Se não houver valores numéricos antes ou depois do bloco, deixar NA
      if (is.na(valor_antes) || is.na(valor_depois)) {
        resultado[bloco_inicio:bloco_fim] <- NA_real_
      } else {
        # Se houver valores antes e depois, preencher com progressão
aritmética
        diferenca <- valor_depois - valor_antes
        tamanho_progressao <- bloco_fim - bloco_inicio + 1
        razao <- diferenca / (tamanho_progressao + 1)
        progressao <- seq(valor_antes + razao, valor_depois - razao,
length.out = tamanho_progressao)
        resultado[bloco_inicio:bloco_fim] <- progressao
      }
    }
  }

  return(resultado)
}

# Aplicar a função para preencher NAs com progressão aritmética por grupo
df_amostra <- df_amostra %>%
  group_by(Imovel) %>%
  mutate(Leitura = preencher_NAs_progressao_aritmetica(Leitura)) %>%
  ungroup()

summary(df_amostra)

# Transformar em data table
df_amostra <- data.table(df_amostra)

# Calcular a diferença entre os valores da leitura em datas consecutivas
df_amostra[, Consumo := Leitura - shift(Leitura), by = Imovel]
df_amostra[, Consumo := replace(Consumo, is.na(Consumo), 0)]

# Retirar valores de consumo negativos
df_amostra <- df_amostra %>%
```

```
  filter(!is.na(Consumo) & Consumo>=0)

# Calcular a média por imóvel
media_imovel <- df_amostra %>%
  group_by(Imovel) %>%
  summarise(Media = mean(Consumo, na.rm = TRUE), .groups = 'drop') %>%
  ungroup()

# Substituir valores de consumo que excedem o limite por NA
df_amostra <- df_amostra %>%
  left_join(media_imovel %>% select(Imovel, Media), by = c("Imovel")) %>%
  mutate(Consumo = ifelse(Consumo > 12.9 * Media, NA, Consumo))

# Extrair apenas o horário do intervalo 4 horas
df_amostra$Intervalo_Hora   <-   format(as.POSIXct(df_amostra$Intervalo_4h),
"%H:%M:%S")

# Calcular média horária de consumo por imóvel e intervalo horário
media_horaria <- df_amostra %>%
  group_by(Imovel, Intervalo_Hora) %>%
  summarise(Media_Horaria = mean(Consumo, na.rm = TRUE), .groups = 'drop')

# Preencher os NAs pelo valor médio horário correspondente
df_amostra <- df_amostra %>%
  left_join(media_horaria, by = c("Imovel", "Intervalo_Hora")) %>%
  mutate(Consumo = ifelse(is.na(Consumo), Media_Horaria, Consumo)) %>%
  select(-Intervalo_Hora, -Media_Horaria)  # Remover colunas intermediárias
se não forem necessárias

# Transformar em data table
df_amostra <- data.table(df_amostra)

# Cálculo de consumo diário
df_diario <- df_amostra %>%
  group_by(Ano = year(Data), Mes = month(Data), Dia = day(Data)) %>%
  summarise(Consumo_Dia = round(mean(Consumo, na.rm = TRUE), 3), .groups =
'drop') %>%
  mutate(Data  =  paste0(Ano,  "-",  ifelse(Mes  <  10,  paste0("0",  Mes),
as.character(Mes)),    "-",    ifelse(Dia    <    10,    paste0("0",    Dia),
as.character(Dia))),
         Data = as.Date(Data))
```

```
# Exportação dos dados de consumo total por dia
df_diario$Consumo_Dia <- gsub("\\.", ",", df_diario$Consumo_Dia)
write.csv2(df_diario, file = "C:/Users/enzod/OneDrive - unb.br/PF2/Dados
Tratados/2022 - Dados Consumo Dia.csv", row.names = FALSE)
```

7.2 Rotina ARIMA e SARIMA no RStudio

```
# Projeto Final 2 - Enzo D'Angelo Arruda Duarte

# CONFIGURAÇÃO DO AMBIENTE
# Limpeza do ambiente
rm(list = ls())

# Carregar pacotes necessários
pacotes <- c("forecast", "tseries", "MASS", "lmtest", "ggplot2",
"data.table", "scales", "beepr", "hydroGOF")
installed_packages <- pacotes %in% installed.packages()
if (any(!installed_packages)) {
  install.packages(pacotes[!installed_packages])
}
invisible(lapply(pacotes, library, character.only = TRUE))

# Definir o diretório de trabalho
setwd("C:/Users/enzod/OneDrive - unb.br/PF2/Dados Tratados")

# Carregar os dados do CSV
df <- fread("2022 - Dados Consumo Dia.csv")

# Limpeza e conversão de dados
df[, Consumo_Dia := as.numeric(gsub(",", ".", Consumo_Dia))]

# Calcular a vazão a partir da coluna Consumo
df[, Vazao := Consumo_Dia / (24) * 1000]

# Verificar e garantir que Data está no formato correto
df$Data <- as.Date(df$Data)

# SÉRIE TEMPORAL
# Criar uma série temporal
TSvazao <- ts(df$Vazao, frequency = 7)
ggtsdisplay(TSvazao)
```

```
TSfinal <- tsclean(TSvazao)
ggtsdisplay(TSfinal)

# Criação do gráfico de análise de outliers
outliers <- tsoutliers(TSvazao)
outliers_df <- data.frame(
  index = time(TSvazao)[outliers$index],
  replacements = outliers$replacements
)

# Criação do gráfico
autoplot(tsclean(TSvazao), series = "Clean", color = '#FF6F61', lwd = 0.1) +
  autolayer(TSvazao, series = "Original", color = 'gray', lwd = 0.1) +
  geom_point(data = outliers_df, aes(x = index, y = replacements), color =
'#004080', size = 2, shape = 21, fill = '#004080') +
  labs(
    x = "Tempo",
    y = "Consumo (L/h)",
    title = "Análise de outliers na série temporal",
    subtitle = "Dados limpos e pontos de outliers destacados"
  ) +
  theme_minimal() +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  theme(
    legend.position = "bottom",
    axis.title.x = element_text(size = 12, margin = margin(t = 10)),
    axis.title.y = element_text(size = 12, margin = margin(r = 10)),
    axis.text = element_text(size = 12, color = "black"),
    plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
    plot.subtitle = element_text( size = 11, hjust = 0.5),
    panel.grid.major = element_line(color = "grey92"),
    panel.grid.minor = element_line(color = "grey92")
  )

# Teste de estacionariedade
adf_result <- adf.test(TSfinal)
print(adf_result)

# Preparar ts_data com Data e Intervalo_4h
ts_data <- data.table(
```

```
  Intervalo_4h = time(TSfinal),  # Manter Intervalo_4h como POSIXct
  Vazao = coredata(TSfinal)
)

# Adicionar uma coluna com números sequenciais para o eixo x
ts_data$Seq <- seq_along(ts_data$Intervalo_4h)

# Calcular o valor médio da vazão
media_vazao <- mean(ts_data$Vazao, na.rm = TRUE)

# Criação do gráfico da série temporal
p <- ggplot(ts_data, aes(x = Seq, y = Vazao)) +
  geom_line(color = "#6BAED6", size = 0.1) +
  geom_hline(yintercept = media_vazao, color = "#FF6F61", linetype =
"dashed", size = 0.5) +
  labs(title = "Série Temporal",
       subtitle = "Dados de consumo médio diário de 2022",
       x = "Tempo",
       y = "Vazão (L/h)") +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5),
    axis.title.x = element_text(size = 12, margin = margin(t = 10)),
    axis.title.y = element_text(size = 12, margin = margin(r = 10)),
    axis.text = element_text(size = 12, color = "black"),
    panel.grid.major = element_line(color = "grey92"),
    panel.grid.minor = element_line(color = "grey92")
  ) +
  scale_x_continuous(breaks = c(1, 32, 60, 91, 121, 152, 182, 213, 244, 274,
305, 335), labels = c("Jan", "Fev", "Mar", "Abr", "Mai", "Jun", "Jul", "Ago",
"Set", "Out", "Nov", "Dez")) +
  scale_y_continuous(labels = scales::comma) +  # Formatando labels do eixo
y com vírgulas
  expand_limits(y = 0)  # Garantir que o valor zero apareça no eixo y

print(p)

# ARIMA
# Dividir os dados em treino (80%) e teste (20%)
tam_treino <- floor(0.8 * length(TSfinal))
```

```
# Criar a série de treinamento
ts_treino <- ts(TSfinal[1:tam_treino])

# Encontrar o início e fim da série de treino
start_treino <- start(ts_treino) # Início da série de treinamento
end_treino <- end(ts_treino)     # Fim da série de treinamento

# Ajustar o inicio da série de teste
start_teste <- c(end_treino[1] + 1)

# Criar a série de teste com o start ajustado
ts_teste <- ts(TSfinal[(tam_treino + 1):length(TSfinal)], start =
start_teste)

# Ajustar o modelo ARIMA
modelo_arima <- auto.arima(ts_treino, stepwise = FALSE)
summary(modelo_arima)

# Teste dos coeficientes do modelo ARIMA
coeftest(modelo_arima)

# Calcular o NSE
nse_arima_ajuste <- NSE(modelo_arima$fitted, ts_treino)
print(nse_arima_ajuste)

# Testes
# Calcular os resíduos para o conjunto de treino ARIMA
residuos_arima <- ts_treino - modelo_arima$fitted

# Verificar os resíduos do modelo ARIMA para o conjunto de treino
checkresiduals(residuos_arima)

# Plotar o histograma dos resíduos
hist(residuos_arima,
     main = "Histograma dos resíduos ARIMA",
     xlab = "Resíduos",
     ylab = "Frequência",
     col = "#6BAED6",
     border = "black",
     prob = TRUE)  # prob = TRUE para obter a densidade ao invés de frequência
```

```
# Adicionar a curva de densidade normal
curve(dnorm(x, mean = mean(residuos_arima), sd = sd(residuos_arima)),
      col = "#004080",
      lwd = 2,
      add = TRUE)

# Plotar o gráfico de dispersão entre os resíduos e os valores previstos
plot(modelo_arima$fitted, residuos_arima,
     main = "Gráfico de dispersão dos resíduos ARIMA",
     xlab = "Vazão prevista (L/h)",
     ylab = "Resíduos",
     pch = 16,
     col = "#6BAED6")
abline(h = 0, col = "#004080", lwd = 2)  # Linha horizontal no eixo zero para
facilitar a visualização

# Criar ts_data_treino com os dados de treinamento
ts_data_treino <- data.table(
  Intervalo_4h = ts_data$Intervalo_4h[1:tam_treino],
  Vazao = ts_treino,
  Seq = ts_data$Seq[1:tam_treino]
)

# Ajustar a coluna fitted no ts_data_treino
ts_data_treino$fitted <- modelo_arima$fitted

# Plotar os dados observados, ajuste ARIMA e dados de treinamento
p <- ggplot() +
  geom_line(data = ts_data, aes(x = Seq, y = Vazao, color = "Dados de
consumo"), size = 0.1) +
  geom_line(data = ts_data_treino, aes(x = Seq, y = fitted, color = "Ajuste
ARIMA"), size = 0.5) +
  labs(title = "Ajuste ARIMA",
       subtitle = "Dados de consumo médios diários de 2022",
       x = "Tempo",
       y = "Vazão (L/h)",
       color = "Legenda") +  # Adiciona título à legenda
  scale_color_manual(values = c("Dados de consumo" = "#6BAED6", "Ajuste
ARIMA" = "#004080")) +
  guides(color = guide_legend(reverse = TRUE)) +  # Inverte a ordem da legenda
```

```
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5),
    plot.caption = element_text(size = 10, hjust = 1),
    axis.title.x = element_text(size = 12, margin = margin(t = 10)),
    axis.title.y = element_text(size = 12, margin = margin(r = 10)),
    axis.text = element_text(size = 12, color = "black"),
    legend.title = element_text(face = "bold", size = 12),
    legend.text = element_text(size = 12),
    panel.grid.major = element_line(color = "grey92"),
    panel.grid.minor = element_blank()
  ) +
  scale_x_continuous(breaks = c(1, 32, 60, 91, 121, 152, 182, 213, 244, 274,
305, 335), labels = c("Jan", "Fev", "Mar", "Abr", "Mai", "Jun", "Jul", "Ago",
"Set", "Out", "Nov", "Dez")) +
  scale_y_continuous(labels = scales::comma) +  # Formatando labels do eixo
y com vírgulas
  expand_limits(y = 0)  # Garantir que o valor zero apareça no eixo y

print(p)

# Teste de acurácia da modelagem
accuracy(modelo_arima$fitted, ts_treino)

# Prever os últimos 20% dos dados
previsao_arima <- forecast(modelo_arima, h = length(ts_teste))

# Teste de acurácia da previsão
accuracy(previsao_arima$mean, ts_teste)

# Calcular o NSE
nse_arima_previsao <- NSE(previsao_arima$mean, ts_teste)
print(nse_arima_previsao)

# Criar ts_data_teste com os dados de teste
ts_data_teste <- data.table(
  Intervalo_4h = ts_data$Intervalo_4h[(tam_treino + 1):nrow(ts_data)],
  Vazao = ts_teste,
  Seq = ts_data$Seq[(tam_treino + 1):nrow(ts_data)],
  Previsao = previsao_arima$mean
```

```
)

# Plotar gráfico de previsão com dados de teste
p <- ggplot() +
  geom_line(data = ts_data, aes(x = Seq, y = Vazao, color = "Dados de
consumo"), size = 0.1) +
  geom_line(data = ts_data_treino, aes(x = Seq, y = fitted, color = "Ajuste
ARIMA"), size = 0.5) +
  geom_line(data = ts_data_teste, aes(x = Seq, y = Previsao, color = "Previsão
ARIMA"), size = 0.5, alpha = 0.8) +
  labs(title = "Previsão ARIMA",
       subtitle = "Comparação entre dados reais e previsões",
       x = "Tempo",
       y = "Vazão (L/h)",
       color = "Legenda") +
  scale_color_manual(values = c("Dados de consumo" = "#6BAED6", "Ajuste
ARIMA" = "#004080", "Previsão ARIMA" = "#FF6F61"),
                     breaks = c("Dados de consumo", "Ajuste ARIMA", "Previsão
ARIMA")) +
  guides(color = guide_legend(reverse = FALSE)) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5),
    plot.caption = element_text(size = 10, hjust = 1),
    axis.title.x = element_text(size = 12, margin = margin(t = 10)),
    axis.title.y = element_text(size = 12, margin = margin(r = 10)),
    axis.text = element_text(size = 12, color = "black"),
    legend.title = element_text(face = "bold", size = 10),
    legend.text = element_text(size = 10),
    panel.grid.major = element_line(color = "grey92"),
    panel.grid.minor = element_blank()
  ) +
  scale_x_continuous(breaks = c(1, 32, 60, 91, 121, 152, 182, 213, 244, 274,
305, 335), labels = c("Jan", "Fev", "Mar", "Abr", "Mai", "Jun", "Jul", "Ago",
"Set", "Out", "Nov", "Dez")) +
  scale_y_continuous(labels = scales::comma) +  # Formatando labels do eixo
y com vírgulas
  expand_limits(y = 0)  # Garantir que o valor zero apareça no eixo y

print(p)
```

```
# SARIMA
# Dividir os dados em treino (80%) e teste (20%)
tam_treino <- floor(0.8 * length(TSfinal))

# Criar a série de treinamento
ts_treino <- ts(TSfinal[1:tam_treino], frequency = 7)

# Criar a série de teste
ts_teste <- ts(TSfinal[(tam_treino + 1):length(TSfinal)], start =
end(ts_treino) + 1, frequency = 7)

# Ajustar o modelo SARIMA
modelo_sarima <- Arima(ts_treino, order =  c(2,0,3), seasonal = c(0,1,3))
modelo_sarima <- auto.arima(ts_treino, seasonal = TRUE, stepwise = FALSE)
summary(modelo_sarima)

# Teste dos coeficientes do modelo SARIMA
coeftest(modelo_sarima)

# Calcular o NSE
nse_sarima_ajuste <- NSE(modelo_sarima$fitted, ts_treino)
print(nse_sarima_ajuste)

# Calcular os resíduos para o conjunto de treino SARIMA
residuos_sarima <- ts_treino - modelo_sarima$fitted

# Verificar os resíduos do modelo SARIMA para o conjunto de treino
checkresiduals(residuos_sarima)

# Plotar o histograma dos resíduos
hist(residuos_sarima,
     main = "Histograma dos resíduos SARIMA",
     xlab = "Resíduos",
     ylab = "Frequência",
     col = "#6BAED6",
     border = "black",
     prob = TRUE)  # prob = TRUE para obter a densidade ao invés de frequência

# Adicionar a curva de densidade normal
curve(dnorm(x, mean = mean(residuos_sarima), sd = sd(residuos_sarima)),
```

```
      col = "#004080",
      lwd = 2,
      add = TRUE)

# Plotar o gráfico de dispersão entre os resíduos e os valores previstos
plot(modelo_sarima$fitted, residuos_sarima,
     main = "Gráfico de dispersão dos resíduos SARIMA",
     xlab = "Vazão prevista (L/h)",
     ylab = "Resíduos",
     pch = 16,
     col = "#6BAED6")
abline(h = 0, col = "#004080", lwd = 2)  # Linha horizontal no eixo zero para
facilitar a visualização

# Previsão com o modelo SARIMA para o conjunto de teste
previsao_sarima <- forecast(modelo_sarima, h = length(ts_teste))

# Teste de acurácia da previsão
accuracy(previsao_sarima$mean, ts_teste)

# Calcular o NSE
nse_sarima_previsao <- NSE(previsao_sarima$mean, ts_teste)
print(nse_sarima_previsao)

# Criar ts_data_treino com os dados de treinamento
ts_data_treino <- data.table(
  Intervalo_4h = ts_data$Intervalo_4h[1:tam_treino],
  Vazao = ts_treino,
  Seq = ts_data$Seq[1:tam_treino],
  fitted = modelo_sarima$fitted
)

# Plotar os dados observados, ajuste SARIMA e dados de treinamento
p <- ggplot() +
  geom_line(data = ts_data, aes(x = Seq, y = Vazao, color = "Dados de
consumo"), size = 0.1) +
  geom_line(data = ts_data_treino, aes(x = Seq, y = fitted, color = "Ajuste
SARIMA"), size = 0.5) +
  labs(title = "Ajuste SARIMA",
       subtitle = "Dados de consumo médios diários de 2022",
       x = "Tempo",
```

```
       y = "Vazão (L/h)",
       color = "Legenda") +  # Adiciona título à legenda
  scale_color_manual(values = c("Dados de consumo" = "#6BAED6", "Ajuste
SARIMA" = "#004080")) +
  guides(color = guide_legend(reverse = TRUE)) +  # Mantém a ordem da legenda
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5),
    plot.caption = element_text(size = 10, hjust = 1),
    axis.title.x = element_text(size = 12, margin = margin(t = 10)),
    axis.title.y = element_text(size = 12, margin = margin(r = 10)),
    axis.text = element_text(size = 12, color = "black"),
    legend.title = element_text(face = "bold", size = 12),
    legend.text = element_text(size = 12),
    panel.grid.major = element_line(color = "grey92"),
    panel.grid.minor = element_blank()
  ) +
  scale_x_continuous(breaks = c(1, 32, 60, 91, 121, 152, 182, 213, 244, 274,
305, 335), labels = c("Jan", "Fev", "Mar", "Abr", "Mai", "Jun", "Jul", "Ago",
"Set", "Out", "Nov", "Dez")) +
  scale_y_continuous(labels = scales::comma) +  # Formatando labels do eixo
y com vírgulas
  expand_limits(y = 0)  # Garantir que o valor zero apareça no eixo y

print(p)

# Criar ts_data_teste com os dados de teste
ts_data_teste <- data.table(
  Intervalo_4h = ts_data$Intervalo_4h[(tam_treino + 1):nrow(ts_data)],
  Vazao = ts_teste,
  Seq = ts_data$Seq[(tam_treino + 1):nrow(ts_data)],
  Previsao = previsao_sarima$mean
)

# Plotar gráfico de previsão com dados de teste
p <- ggplot() +
  geom_line(data = ts_data, aes(x = Seq, y = Vazao, color = "Dados de
consumo"), size = 0.1) +
  geom_line(data = ts_data_treino, aes(x = Seq, y = fitted, color = "Ajuste
SARIMA"), size = 0.5) +
```

```
  geom_line(data = ts_data_teste, aes(x = Seq, y = Previsao, color = "Previsão
SARIMA"), size = 0.5, alpha = 0.8) +
  labs(title = "Previsão SARIMA",
       subtitle = "Comparação entre dados reais e previsões",
       x = "Tempo",
       y = "Vazão (L/h)",
       color = "Legenda") +
  scale_color_manual(values = c("Dados de consumo" = "#6BAED6", "Ajuste
SARIMA" = "#004080", "Previsão SARIMA" = "#FF6F61"),
                     breaks = c("Dados de consumo", "Ajuste SARIMA",
"Previsão SARIMA")) +
  guides(color = guide_legend(reverse = FALSE)) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5),
    plot.caption = element_text(size = 10, hjust = 1),
    axis.title.x = element_text(size = 12, margin = margin(t = 10)),
    axis.title.y = element_text(size = 12, margin = margin(r = 10)),
    axis.text = element_text(size = 12, color = "black"),
    legend.title = element_text(face = "bold", size = 10),
    legend.text = element_text(size = 10),
    panel.grid.major = element_line(color = "grey92"),
    panel.grid.minor = element_blank()
  ) +
  scale_x_continuous(breaks = c(1, 32, 60, 91, 121, 152, 182, 213, 244, 274,
305, 335), labels = c("Jan", "Fev", "Mar", "Abr", "Mai", "Jun", "Jul", "Ago",
"Set", "Out", "Nov", "Dez")) +
  scale_y_continuous(labels = scales::comma) +  # Formatando labels do eixo
y com vírgulas
  expand_limits(y = 0)  # Garantir que o valor zero apareça no eixo y

print(p)
```

7.3 Rotina Regressão Linear Multivariada no RStudio

```
# Projeto Final 2 - Enzo D'Angelo Arruda Duarte

# CONFIGURAÇÃO DO AMBIENTE
# Limpeza do ambiente
rm(list = ls())
```

```
# Carregar pacotes necessários
pacotes <- c("dplyr", "car", "rstatix", "lmtest", "ggpubr", "psych",
"lm.beta", "data.table","hydroGOF")
installed_packages <- pacotes %in% installed.packages()
if (any(!installed_packages)) {
  install.packages(pacotes[!installed_packages])
}
invisible(lapply(pacotes, library, character.only = TRUE))

# Definir o diretório de trabalho
setwd("C:/Users/enzod/OneDrive - unb.br/PF2/Dados Tratados")

# Carregar os dados do CSV
df <- fread("2022 - Dados Consumo Dia.csv")

# Limpeza e conversão de dados
df[, Consumo_Dia := as.numeric(gsub(",", ".", Consumo_Dia))]
df[, Data := as.POSIXct(Data, format = "%Y-%m-%d")]

# Calcular a vazão a partir da coluna Consumo
df[, Vazao := Consumo_Dia / (24) * 1000]

# Criar uma nova coluna para identificar o dia da semana
df$Dia_Semana <- weekdays(df$Data)

# Criar uma nova coluna para a vazão média iniciada no dia
df <- df %>%
  arrange(Data) %>%
  mutate(Qm = sapply(1:nrow(.), function(i) {
    start_index <- i
    end_index <- min(i + 7 - 1, nrow(.))
    mean(Vazao[start_index:end_index], na.rm = TRUE)
  }))

# Visualizar os primeiros registros para verificar
head(df)

# Criar a coluna "Q_dia_anterior" (Vazão do dia anterior)
df <- df %>%
  arrange(Data) %>%
  mutate(Q_dia_anterior = lag(Vazao, n = 1, default = 0))
```

```
# Adicionar colunas para os valores de vazão do mesmo dia da semana em semanas
anteriores
df <- df %>%
  mutate(Q_semana_1 = lag(Vazao, n = 7, default = 0),
         Q_semana_2 = lag(Vazao, n = 14, default = 0),
         Q_semana_3 = lag(Vazao, n = 21, default = 0))

# Remover as colunas 1, 2, 3, 4, 5 e 7
df <- df %>%
  select(-c(1, 2, 3, 4, 5, 7))

# REGRESSÃO LINEAR MULTIVARIADA
# Modelo
modelo_rlm <- lm(Vazao ~ Qm + Q_dia_anterior + Q_semana_1 + Q_semana_2 +
Q_semana_3, data = df)

# Testes
# Ausência de multicolinearidade
pairs.panels(df)

# Análise
summary(modelo_rlm)
AIC(modelo_rlm)
```